

MSC

2.º
CICLO

FCUP
2022

U.PORTO

When pathways cross other pathways, that themselves, have met other pathways: crabs responses to cocktails of contaminants

Sara Sousa Moreira

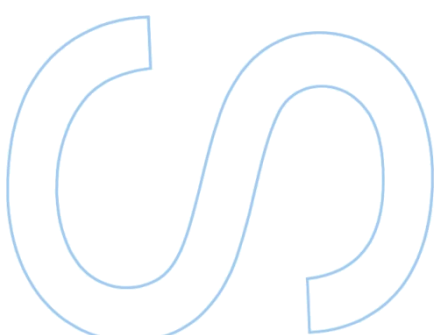
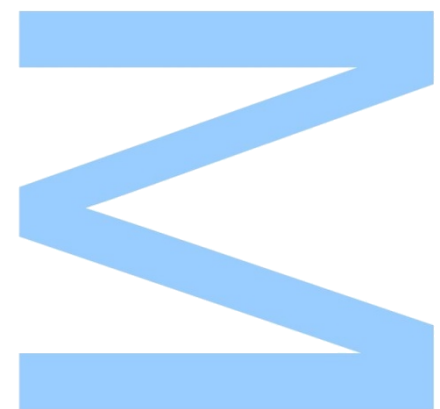
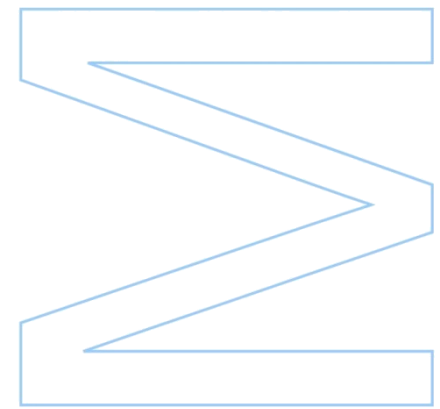
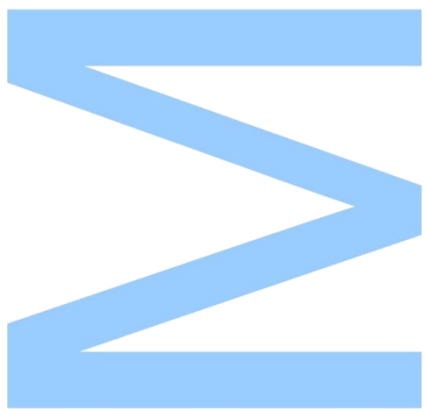
FC



When pathways cross other pathways, that themselves, have met other pathways: crabs responses to cocktails of contaminants

Sara Sousa Moreira

Dissertação de Mestrado apresentada à
Faculdade de Ciências da Universidade do Porto em
Bioinformática e Biologia Computacional
2022



When pathways cross other pathways, that themselves, have met other pathways: crabs responses to cocktails of contaminants

Sara Sousa Moreira

Dissertação de Mestrado apresentada à
Faculdade de Ciências da Universidade do
Porto em Bioinformática e Biologia
Computacional
2022

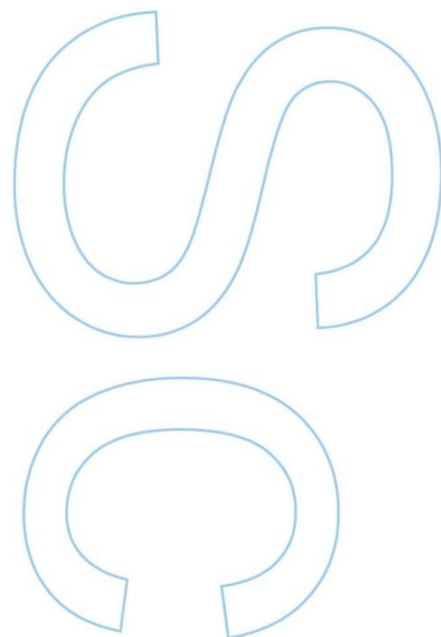
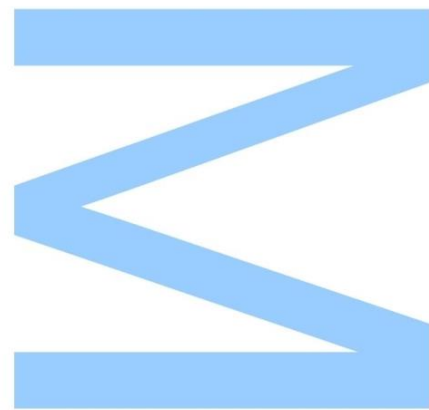
Orientador

Nuno Ferreira, Centro Interdisciplinar de Investigação
Marinha e Ambiental (CIIMAR) e School of Biosciences - Cardiff University

Coorientadores

Laura Guimarães, Centro Interdisciplinar de Investigação
Marinha e Ambiental (CIIMAR)

Pedro G. Ferreira, Departamento de Ciências de Computadores,
FCUP

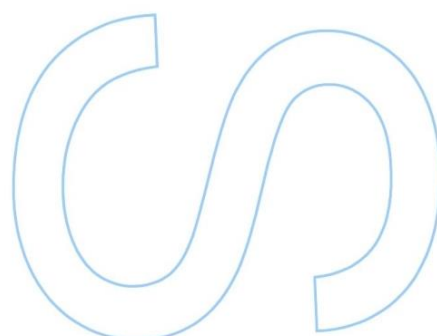
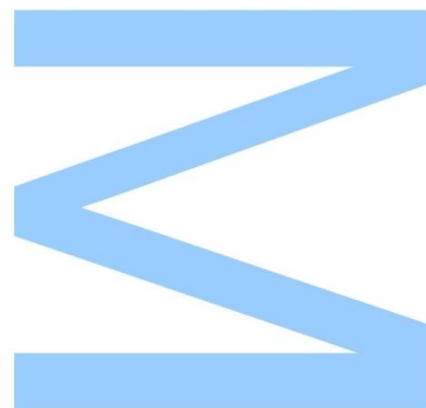




Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, ____/____/____



Declaração de Honra

Eu, Sara Sousa Moreira, inscrito(a) no Mestrado em Bioinformática e Biologia Computacional da Faculdade de Ciências da Universidade do Porto declaro, nos termos do disposto na alínea a) do artigo 14.º do Código Ético de Conduta Académica da U.Porto, que o conteúdo da presente dissertação/relatório de estágio/projeto “When pathways cross other pathways, that themselves, have met other pathways: Crabs responses to cocktails of contaminants” reflete as perspetivas, o trabalho de investigação e as minhas interpretações no momento da sua entrega.

Ao entregar esta dissertação/relatório de estágio/projeto “When pathways cross other pathways, that themselves, have met other pathways: Crabs responses to cocktails of contaminants”, declaro, ainda, que a mesma é resultado do meu próprio trabalho de investigação e contém contributos que não foram utilizados previamente noutros trabalhos apresentados a esta ou outra instituição.

Mais declaro que todas as referências a outros autores respeitam escrupulosamente as regras da atribuição, encontrando-se devidamente citadas no corpo do texto e identificadas na secção de referências bibliográficas. Não são divulgados na presente dissertação/relatório de estágio/projeto “When pathways cross other pathways, that themselves, have met other pathways: Crabs responses to cocktails of contaminants” quaisquer conteúdos cuja reprodução esteja vedada por direitos de autor.

Tenho consciência de que a prática de plágio e auto-plágio constitui um ilícito académico.

Sara Sousa Moreira

Porto, 30 de Setembro de 2022

Agradecimentos

Um grande obrigado ao meu orientador, Dr. Nuno Ferreira. A tua paciência, apoio e prontidão em ajudar foram indispensáveis para conseguir concluir este trabalho. Estar-te-ei sempre grata. Obrigada por me acompanhares neste desafio.

Agradeço igualmente à minha coorientadora, Dra. Laura Guimarães, por se disponibilizar em me esclarecer e ajudar, e ser sempre compreensiva para comigo.

Um obrigado ao meu coorientador, professor Pedro Ferreira. Agradeço o esforço por erguer e manter este Mestrado, sempre com qualidade. Agradeço igualmente a sua ajuda e acessibilidade.

Agradeço igualmente ao Professor Peter Kille pela sua participação e ajuda nas correções finais do meu projeto de Dissertação.

Um muito especial obrigado ao meu namorado e à minha irmã. Sem vocês teria sido muito difícil ultrapassar este desafio. Foram o meu maior apoio nos momentos menos bons e a minha fonte de motivação diária. Obrigada por me ajudarem a concluir mais uma etapa na minha vida.

Um obrigado à minha família.

Por fim, um muito obrigado à Dra. Ana Nobre.

Abstract

Intense anthropogenic occupation near estuaries, and consequent increased pollutants discharges, have compromised wild aquatic organisms' survival and ecosystem functioning. These transitional zones harbour several species highly subjected to natural stressors and daily compound exposure. Transcriptomic responses are one of the most accurate biomarkers to determine the impact of stressors at lower organizational levels, providing the main genes and metabolic pathways being affected when exposed to an unknown or very complex mixture of contaminants. Their use has been particularly successful in the often-neglected crustaceans.

In the present study, the green crab *Carcinus maenas* was selected as a model organism to investigate possible differential transcriptomic responses when exposed to environments with distinct levels of contamination. The assessments were possible due to bioinformatic analyses of high-throughput transcriptomic data. The crabs were previously collected in the north-western Portuguese estuaries of the Minho and Lima rivers, differing in human occupation and levels of contamination. The Lima river is under moderate anthropogenic pressure, and the Minho river is under a low anthropogenic pressure. These macroinvertebrates have shown to be an adequate experimental model to investigate the effects of environmental contaminants at a transcriptomic level and assess the status of estuarine ecosystems.

Multiple computational softwares were applied in the present thesis, allowing the extracted RNA to be exploited. A preliminary analysis using FastQC and Trimmomatic tools reported the quality of each extracted sample and removed poor quality reads and adapters inserted for the molecular sequencing, respectively. The results obtained with the assessment of quality using FastQC suggested a biased library and problems in the sequencing process. The trimming step was insufficient to improve the quality of the samples' sequences. Since the *C. maenas* transcriptome was not sequenced and assembled at the beginning of this study, a *de novo* assembly was conducted upon the trimmed sequences of each collected sample, using the Trinity assembler. The reconstructed transcripts' quality was first assessed through primary statistics, complemented with three computational analyses: a Principal Component Analysis, a MA plot, and a Dendrogram and Heatmap. The statistics provided to evaluate the overall quality of the reconstructed assemblies supported the predominance of discrepancy. The three computational analyses suggested minimal differences between the samples from both estuaries. In parallel, the same assembly strategy was conducted upon all raw sequences of each sample concatenated. The objective of

performing a second analysis was to verify if merging all raw reads prior to a *de novo* assembly would lead to a better reconstructed transcript. In the present work, the process of concatenation appears to have improved the final assembled transcript. Given the occurrence of highly similar transcripts when assembling a transcriptome, the EvidentialGene software was used to remove duplicates and redundancy. In the present study, low-quality sequences increased this similarity. Besides conducting a redundancy assessment upon each sample's transcriptome, the removal of redundant sequences was also performed upon all assemblies concatenated and upon the transcript reconstructed from the prior concatenated raw reads. Once again these analysis were performed to determine whether the process of concatenation would result in a better reconstructed transcript. However, it was not possible to determine with certainty whether the concatenation process improves or not the final reconstructed transcriptome. Lastly, the assembly completeness was assessed using BUSCO, providing percentages of genes under single-copy control for each transcript. The process of concatenation improved the completeness results, particularly when assembling a transcriptome from a set of concatenated raw reads. A superior percentage of genes under single-copy control and a much more reduced number of missing genes were observed.

The present study looks to enhance the Crustacea sub-phylum, particularly the green crab, as suitable for ecotoxicological scientific research. This macroinvertebrate is one of the best-studied estuarine species, demonstrates sensitivity to abiotic and chemical stress and is an important biological indicator of the impact of environmental contamination. Moreover, it highlights the use of bioinformatic and computational approaches as adequate to analyse high-throughput transcriptomic data. Despite the errors in the sequencing process carried out in previous research, the computational softwares applied along this thesis allowed to analyze sequences of RNA.

Keywords: Bioinformatic; High-throughput transcriptomic data; Computational tools; *Carcinus maenas*; Estuarine anthropogenic contamination; Differential transcriptomic responses.

Resumo

A intensa ocupação antropogénica próxima aos estuários, e consequente aumento da descarga de poluentes, tem comprometido a sobrevivência dos organismos aquáticos selvagens e o funcionamento dos ecossistemas. Estas zonas de transição servem de abrigo a várias espécies, altamente sujeitas a *stressors* naturais, e exposição diária a compostos nocivos. As respostas transcriptómicas são um dos biomarcadores mais precisos para determinar o impacto de *stressors* a níveis organizacionais mais baixos, fornecendo os principais genes e vias metabólicas afetados, quando expostos a uma mistura desconhecida ou muito complexa de contaminantes. O seu uso tem sido particularmente bem-sucedido nos crustáceos, muitas vezes negligenciados.

No presente estudo, o caranguejo verde *Carcinus maenas* foi selecionado como o organismo modelo para investigar possíveis respostas transcriptómicas diferenciais, quando exposto a ambientes com níveis distintos de contaminação. As avaliações foram possíveis devido a análises bioinformáticas de “grandes dados” transcriptómicos. Os caranguejos foram previamente recolhidos nos estuários portugueses dos rios Minho e Lima, que diferem na ocupação humana e nos níveis de contaminação: o rio Lima está sob pressão antropogénica moderada, e o rio Minho está sob pressão antropogénica reduzida. Estes macroinvertebrados demonstraram ser um modelo experimental adequado para investigar os efeitos de contaminantes ambientais ao nível transcriptómico, e avaliar os ecossistemas estuarinos.

Vários softwares computacionais foram aplicados ao longo da presente tese, permitindo explorar as sequências de RNA extraídas. A análise preliminar realizada com a utilização das ferramentas FastQC e Trimmomatic relatou a qualidade de cada amostra extraída e removeu *reads* de baixa qualidade e adaptadores inseridos para a sequenciação molecular, respetivamente. Os resultados obtidos pelo FastQC sugeriram erros durante a construção da biblioteca e problemas no processo de sequenciação. A etapa de *trimming* foi insuficiente para melhorar a qualidade dos sets de sequências das amostras. Dado que o transcriptoma da espécie *C. maenas* não se encontrava ainda sequenciado e reconstruído no princípio deste estudo, um *assembly de novo* foi realizado sob as *trimmed reads* com o *assembler* Trinity. A qualidade dos transcritos reconstruídos foi primeiramente avaliada através de estatísticas primárias, complementadas com três análises computacionais: *Principal Component Analysis*, gráfico MA, Dendrograma e *Heatmap*. As estatísticas fornecidas para avaliar a

qualidade geral dos assemblies reconstruídos apoiaram a predominância de discrepância. As três análises computacionais sugeriram diferenças mínimas entre as amostras de ambos os estuários. Paralelamente, a estratégia de assembly *de novo* foi realizada sob todas as *raw reads* concatenadas. O objetivo de realizar uma segunda análise foi verificar se o processo de concatenação resultaria num transcrito final melhor reconstruído. No presente estudo, o processo de concatenação parece ter melhorado o transcriptoma final. Dada a ocorrência de transcritos altamente semelhantes durante a reconstrução de um transcriptoma, o software EvidentialGene foi utilizado para remover redundância e sequências duplicadas. No presente estudo, sequências com baixa qualidade aumentaram a similaridade entre transcritos. Para além de se realizar uma avaliação de redundância sob o transcriptoma de cada amostra, a remoção de sequências redundantes também foi aplicada sob todos os assemblies concatenados e sob o transcrito reconstruído a partir das *raw reads* concatenadas. Novamente, estas análises foram realizadas com o objetivo de determinar se o processo de concatenação resultaria num transcrito melhor reconstruído. No entanto, não foi possível estabelecer com certeza se o processo de concatenação melhora ou não o transcriptoma final reconstruído. Por fim, foi realizada uma avaliação de *completeness* de cada assembly com a ferramenta BUSCO. Este programa forneceu percentagens de *single-copy genes* e *missing genes* para cada transcrito. O processo de concatenação melhorou os resultados de *completeness*, particularmente quando um transcriptoma é reconstruído a partir de *raw reads* concatenadas. Uma percentagem superior de *single-copy genes* e um número muito mais reduzido de *missing genes* foram observados.

A presente tese procurou evidenciar o subfilo Crustacea, particularmente o caranguejo verde (*C. maenas*), como adequado para estudos científicos ecotoxicológicos. Este macroinvertebrado é uma das espécies estuarinas mais bem estudadas, demonstra sensibilidade a stress abiótico e químico, e é um importante indicador biológico do impacto da contaminação ambiental. Além disso, evidencia o uso da bioinformática e abordagens computacionais como adequadas para analisar dados transcriptômicos. Apesar dos erros no processo de sequenciação realizado em pesquisas anteriores, os softwares computacionais aplicados ao longo desta tese permitiram analisar sequências de RNA.

Palavras-chave: Bioinformática; Análise de “grandes dados” transcriptômicos; Ferramentas computacionais; *Carcinus maenas*; Contaminação antropogénica estuarina; Respostas transcriptómicas diferenciais.

Contents

Declaração de Honra	5
Agradecimentos	7
Abstract	9
Resumo	12
Contents	15
List of Figures	19
List of Supplementary Figures	21
List of Tables	22
List of Abbreviations	24
Chapter I. Introduction	28
1. Model species and sampling locations	28
1.1. The green crab <i>Carcinus maenas</i>	28
1.1.1. Morphology	29
1.1.2. Life Cycle	31
1.1.3. Endocrine system and moult cycle	31
1.1.4. Model organism in ecotoxicological studies	34
1.2. Estuaries of Minho and Lima Rivers	35
2. Sequencing	38
2.1. Initial advances in sequencing technologies	38
2.2. High-throughput Next Generation Sequencing methods	39
2.3. High-throughput Next-Generation Sequencing platforms: second and third generations	41
2.4. Illumina Sequencer	44
2.5. High-throughput sequencing quality assessment	47
3. Thesis objectives	50
Chapter II. Materials and Methods	52
1. Workflow I	52
1.1. Crab collection and estuaries sampling sites	52
1.2. RNA extraction	52
1.3. Library construction and sequencing	52
2. Workflow II	53
2.1. Sequencing quality control using FastQC	53
2.2. Read trimming using Trimmomatic	54

2.2.1. Implementation of trimming steps	55
2.3. MobaXterm Terminal	58
2.4. <i>de novo</i> Transcriptome Assembly using Trinity.....	58
2.4.1 Transcriptome assembly approaches.....	59
2.4.1.1. Reference-based strategy.....	59
2.4.1.2. <i>de novo</i> assembly strategy.....	61
2.4.1.2.1. Overlap-layout-consensus approach.....	62
2.4.1.2.2. de Bruijn graph concept	63
2.4.2. Trinity: a method for <i>de novo</i> transcriptome assembly	64
2.4.2.1. Trinity implementation and requirements for assembly.....	66
2.4.3. Reconstruction of eight <i>de novo</i> transcriptome assemblies.....	67
2.4.4. Reconstruction of a single <i>de novo</i> transcriptome assembly	68
2.4.5. Assess the overall quality of the assemblies	69
2.4.5.1. Computational tools for data analysis and visualization.....	72
2.5. Assessment of Redundancy and Completeness.....	74
2.5.1. Evaluation and removal of redundancy	76
2.5.1.1. EvidentialGene: a tool for transcript and gene sets reconstruction and redundancy removal	77
2.5.1.1.1. EvidentialGene pipeline: tr2aacds	78
2.5.1.2. Analysis 1.1. Eight <i>de novo</i> transcriptome assemblies	79
2.5.1.3. Analysis 1.2. <i>de novo</i> transcriptome assemblies concatenated	80
2.5.1.4. Analysis 2. Single reconstructed assembly	81
2.5.1.5. Metrics to evaluate the overall quality of all Evigene analyses	81
2.5.2. Assessment of Completeness.....	82
2.5.2.1. BUSCO: a tool for completeness assessment.....	83
2.5.2.1.1. Setup and implementation of BUSCO software	86
2.5.2.2. Completeness assessment using BUSCO tool.....	87
2.5.2.3. Pre-Evigene metrics for an overall quality comparison	88
Chapter III. Results and Discussion	90
1. Sequencing quality assessment.....	90
2. Transcriptome sequencing output and Trimming	94
3. <i>de novo</i> Transcriptome Assemblies output	96
3.1. Supplementary computational analysis	101
4. Assemblies redundancy output (and pre-evigene comparison)	106
5. Assemblies completeness output.....	109
6. Overall discussion of the results.....	114

Chapter IV. Conclusions	117
1. Final Remarks.....	117
2. Future Work.....	119
Appendix.....	121
Bibliography.....	128

List of Figures

Figure 1 – Green crab <i>Carcinus maenas</i>	28
Figure 2 – Abdomen structure of <i>C. maenas</i> males and females.....	29
Figure 3 – Differential colouration of <i>C.maenas</i> abdomen.....	30
Figure 4 – Geographic location of Minho and Lima rivers.....	37
Figure 5 – Disposition of Lima river estuary.....	38
Figure 6 – Second-generation HT-NGS platforms leading the sequencing field.....	42
Figure 7 - Third-generation HT-NGS platforms leading the sequencing field.....	43
Figure 8 – Sequencing-by-synthesis approach used by Illumina sequencer.....	45
Figure 9 - Association between computation and analysis of biological data.....	51
Figure 10 - Representation of both single-end and paired-end mode strategies.....	55
Figure 11 - Overview of the Trimmomatic script, and respective trimming steps.....	57
Figure 12 - <i>de novo</i> Assembly approach step.....	59
Figure 13 - de Bruijn graph approach.....	64
Figure 14 – Overview of Trinity modules.....	65
Figure 15 – Overview of the Trinity assembler script for Analysis 1.....	67
Figure 16 - Overview of the Trinity assembler script for Analysis 2.....	69
Figure 17 – Overview of the Trinity statistics script for Analysis 1.....	69
Figure 18 - Overview of the Trinity statistics script for Analysis 2.....	70
Figure 19 - Statistical computation and assessment steps for Analyses 1 and 2.....	75
Figure 20 - EvidentialGene script, implementing the tr2aacds pipeline, for Analysis 1.1.....	79
Figure 21 - EvidentialGene script, implementing the tr2aacds pipeline, for Analysis 1.2.....	80

Figure 22 - EvidentialGene script, implementing the tr2aacds pipeline, for Analysis 2.....	81
Figure 23 - BUSCO assessment workflows.....	84
Figure 24 - Total read count per reconstructed assembly (Analyses 1 and 2).....	97
Figure 25 – Total assembled transcripts per reconstructed assembly (Analyses 1 and 2).....	98
Figure 26 – Visualization of two-dimensional PCA plot.....	102
Figure 27 – Visualization of two-dimensional MA plot.....	103
Figure 28 – Visualization of Dendrogram and Heatmap plot.....	105

List of Supplementary Figures

Supplementary Figure 1 – FastQC ‘Per Base Sequence Content’ module plot, for sample C10.....	122
Supplementary Figure 2 - FastQC ‘Per Base Sequence Content’ module plot, for sample C13.....	123
Supplementary Figure 3 – FastQC ‘Per Sequence GC Content’ module plot, for sample C10.....	124
Supplementary Figure 4 – FastQC ‘Per Sequence GC Content’ module plot, for sample C13.....	125
Supplementary Figure 5 – FastQC ‘Sequence Duplication Levels’ plot, for sample C10.....	126
Supplementary Figure 6 – FastQC ‘Sequence Duplication Levels’ plot, for sample C13.....	127

List of Tables

Table 1 – Summarized evaluation of *C.maenas* samples using FastQC tool.....90

Table 2 - Transcriptome sequencing and Trimming results per sample.....95

Table 3 - Trinity quality evaluator metrics.....100

Table 4 - EvidentialGene quality assessment metrics.....108

Table 5 - BUSCO assessment metrics, with no removal of redundant transcripts.....112

Table 6 – BUSCO assessment metrics, with removal of redundant transcripts.....113

List of Abbreviations

- (XO/SG)** X-organ-sinus gland
- (YOs)** Y-organs
- (SG)** Sinus Gland
- (XO)** X-organ
- (GIH)** Gonad Inhibiting Hormone
- (MOIH)** Mandibular Organ Inhibiting Hormone
- (MIH)** Moulting Inhibiting Hormone
- (NDH)** Neurodepressing Hormone
- (CHH)** Crustacean Hyperglycemic Hormone
- (ITP)** Ion Transport Peptide
- (VIH)** Vitellogenesis Inhibiting Hormone
- (AGH)** Androgenic Gland
- (MF)** Methyl Farnesoate
- (MO)** Mandibular Organ
- (JH III)** Juvenile Hormone III
- (Cu)** Copper
- (Cd)** Cadmium
- (Co)** Cobalt
- (Zn)** Zinc
- (PCBs)** Polychlorinated Biphenyls
- (PAHs)** Polycyclic Aromatic Hydrocarbons
- (dNTPs)** Deoxy-nucleotides
- (ddNTPs)** Dideoxy-nucleotides
- (HTS)** High-Throughput Sequencing

(NGS) Next-Generation Sequencing

(HT-NGS) High-Throughput Next-Generation Sequencing

(cDNA) Complementary DNA

(RNA-Seq) RNA Sequencing

(SMRT) Single Molecule Real-Time

(RIN) RNA Integrity Number

(tRNA) Transfer RNA

(mRNA) Messenger RNA

(rRNA) Ribosomal RNA

(QC) Quality Control

(SE) Single-End

(PE) Paired-End

(SSH) Security Shell

(FQ) FastQ

(.pl) Perl Script

(PCA) Principal Component Analysis

(PCs) Principal Components

(SVD) Singular Value Decomposition

(DEGs) Differentially Expressed Genes

(CDS) Coding Sequences

(BUSCO) Benchmarking Universal Single-Copy Orthologues

(OrthoDB) Ortholog Catalogue

(HMMs) Hidden Markov Models

(ORF) Opening Reading Frame

(HPC) High Performance Computing

(FDR) False Discovery Rate

(AChE) Acetylcholinesterase

(GST) Glutathione-S-transferases

(GSH) Glutathione

(NAGase) N-acetyl-b-glucosaminidase

(LDH) Lactate Dehydrogenase

(ICDHs) NADP⁺-dependent isocitrate dehydrogenase

(CAT) Catalase

(TG) Total Glutathione

(SOD) Superoxide Dismutase

(MTs) Metallothioneins

Chapter I. Introduction

1. Model species and sampling locations

1.1. The green crab *Carcinus maenas*

Carcinus maenas (Figure 1) is a key crustacean commonly found along estuarine and coastal ecosystems, previously described by Linneaus in 1758, belonging to the Class Malacostraca, Order Decapoda and Family Portunidae. It is recognised by different designations such as green crab, European green crab and European shore crab.

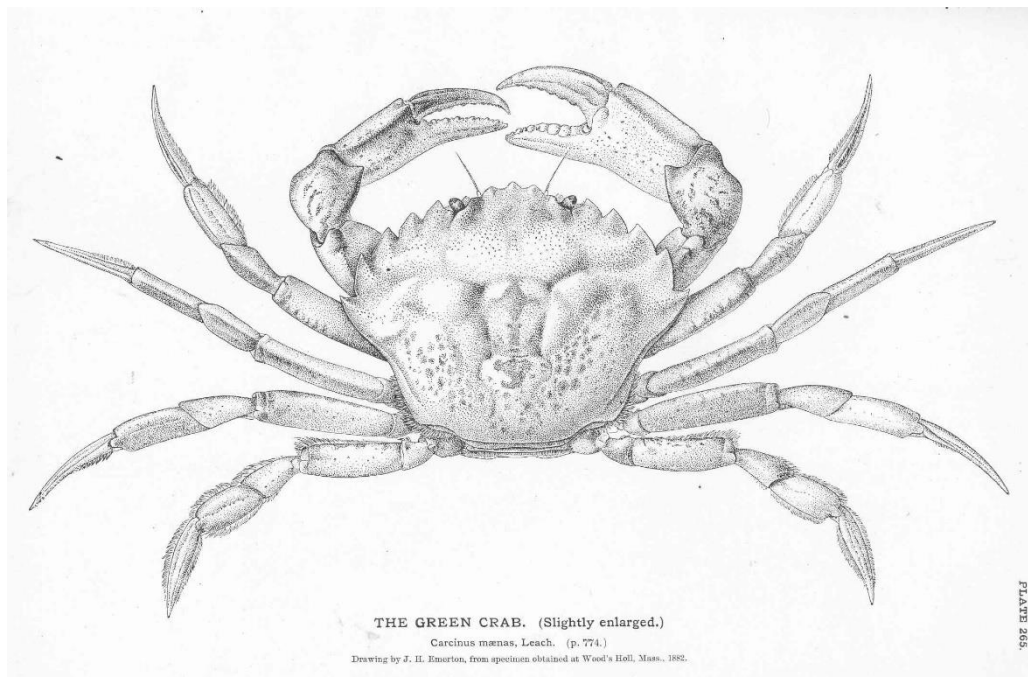


Figure 1 – Green crab *Carcinus maenas*. Image retrieved from (Goode 1884).

Carcinus maenas is native to the European and North African coasts. However, it successfully spread and colonised several worldwide regions, for instance, North and South America, Australia, South Africa, Canada and East Asia. It inhabits rocky and unvegetated intertidal areas, mud and sand, seagrasses and saltmarshes, preferably in protected coastlines with recesses forming a bay, known as embayment (Ray 2005; Grosholz and Ruiz 1996). This epibenthic invertebrate is an omnivorous predator

feeding on detritus, algae, molluscs, snails, annelids, bivalves and other benthic organisms, presenting diet changes as the crabs age (ROPEZ 1968).

1.1.1. Morphology

Externally, the green crab is constituted by a segmented thorax and a symmetrical but reduced abdomen located underneath it. It possesses a well-developed and flattened carapace that dorsoventrally covers the thorax and the lateral gills (Crothers 1967). The carapace width can be used to differentiate and identify crabs sex (Ledesma, Van der Molen, and Barón 2010). Males carapace breadth (up to 86 mm) is usually bigger compared to females carapace (up to 70 mm). The abdomen characteristics can equally facilitate gender identification. Males exhibit triangular abdomens with five segments, while females present rounded and broad abdomens with seven segments (Crothers 1967). Figure 2 illustrates both male and female abdomens.



Figure 2 – *Carcinus maenas* female (left) and male (right) abdomen, respectively. Images retrieved from https://www.wikiwand.com/pt/Carcinus_maenas website.

The decapod Crustacea present an integument comprised of a rigid exoskeleton or cuticle as a primary structure. It covers all body parts and comprises two main layers: the epicuticle and the procuticle. The thin outer layer, known as epicuticle, appears not to endue chitin and consists of a thin outer membrane and an inner thicker layer (Roer and Dillaman 1984). The procuticle is divided into an external exocuticle hardened by calcification and composed of chitin, along with an internal endocuticle heavily calcified

(Travis 1955; Travis 1965). Underneath, a membranous and not calcified layer contacts the epidermis responsible for the cuticle secretion (Travis 1955; Travis 1965). During premolt and ecdysis, the cuticle goes through a process of digestion and absorption of its old structure, and a new hardened and expanded one is secreted.

Adult *C. maenas* exhibit a wide variation of carapace and integument colour from pale green through orange to red. Green colour develops immediately after moulting while red colouration enhances through intermoult (Kaiser, Hughes, and Reid 1990). This colour variation is illustrated in Figure 3. The likelihood of red colour settlement on the integument increases with a longer intermoult phase (Reid et al. 1997). Carapace colour seems to have a connection with behavioural and physiological differentiation (Reid and Aldrich 1989; McGaw et al. 1992). Green crabs are efficient osmoregulators, presenting an higher tolerance to salinity fluctuations and aerial exposure, thus predominantly inhabiting the intertidal zone and salt marshes (McGaw et al. 1992; Reid et al. 1997). Red crabs, prevalent in the subtidal zone, are aggressive dominants given their superior muscular strength and structurally stronger carapaces (Kaiser, Hughes, and Reid 1990), possessing a competitive advantage during fights and increased mating success (Reid et al. 1997).



Figure 3 – *Carcinus maenas* male abdomen in red (left) and green (right) colours. Images retrieved from https://www.wikiwand.com/pt/Carcinus_maenas website.

Carcinus maenas presents multiple segmented appendages arranged with one pair to each embryonic segment (Crothers 1967). Females' and males' thorax present maxillipeds (food manipulation) and five pairs of pereopods (locomotion segments). The first pair presents a claw structure with chelae. The antennae (sensory segments), mandibles and maxillae (feeding segments), and maxillipeds (food manipulation) are

located in the anterior part of the cephalothorax. Underneath the abdomen are located the reproductive appendices (pleopods), presenting sexual dimorphism.

1.1.2. Life Cycle

Green crabs' life cycle alternates between benthic adult and planktonic larval stages with limited swimming capabilities (Klassen and Locke 2007; Chia, Buckland-Nicks, and Young 1984). The pelagic larval stage comprises four zoeal and a megalopal larval stage with dependent growth and development upon food availability, salinity and temperature levels (dos Santos et al. 2008; Leignel et al. 2014). Matting can occur multiple times within a year, with an annual production of one or two clutches of eggs (Klassen and Locke 2007). Still, pre-moult females release pheromones attracting males that carry them for several days until moulting occurs (Klassen and Locke 2007). Mating occurs after female moulting. After internal fecundation, eggs hatch into free-swimming zoea, enclosed in a thin embryonic membrane, at the mouth of the estuary (Crothers 1967; Queiroga, Costlow, and Moreira 1997). Zoea vertical migration patterns enhance their export from estuaries to adjacent shelf waters. Posteriorly, older zoeal stages develop into the post larval stage, megalopa (Queiroga, Costlow, and Moreira 1997; Zeng and Naylor 1996). The final metamorphosis originates the juvenile benthic crab (Crothers 1967). Sexual maturity is attained by three years of age, with males possessing only two pairs of pleopods for copulation and females enduing more individualized reproductive pairs, mainly used for egg carrying and incubation (Crothers 1967).

1.1.3. Endocrine system and moult cycle

Decapod crustaceans, a main taxon of the arthropods phylum, developed endocrine systems with unique characteristics through evolution, highlighted by the development of true endocrine glands with origin in epithelial tissues (DeFur 2004). The appearance of a well-developed and rigid exoskeleton deepened the arthropods' endocrine system divergence. Moulting is required for the growth and reproduction of the exoskeleton, regulated by endocrine signalling processes (Knigge, LeBlanc, and Ford 2021). Crustaceans' neuroendocrine regulation is comprised of an X-organ-sinus gland (XO/SG) complex located in the eyestalk ganglia, and epithelioid glands, referred to as Y-organs (YOs), located in the maxilla (Hopkins 2012).

The sinus gland (SG) is a cluster with a bulbous format, with neurosecretory axonal terminals lying between the external medulla and internal medulla of the eyestalk ganglia, in close contact with the haemolymph (Hopkins 2012). It is a neurohemal storage organ whose axonal endings belong to cell bodies located in a portion of the medulla terminalis known as the X-organ (XO). This organ is responsible for the production of neuropeptide hormones, partially controlled by the neuromodulators dopamine and serotonin (Carlisle and Passano 1953; Rodríguez, Medesani, and Fingerman 2007; Fingerman et al. 1974). Under the appropriate stimulus, the sinus gland secretes several stored peptide hormones into the circulation, including hormones targeting secondary endocrine glands - GIH (gonad inhibiting hormone), MOIH (mandibular organ inhibiting hormone), MIH (moult inhibiting hormone) - and hormones targeting somatic tissues - NDH (neurodepressing hormone) and CHH, referred to as crustacean hyperglycemic hormone (Rodríguez, Medesani, and Fingerman 2007).

The first CHH to be fully sequenced and characterised was isolated from the green crab *Carcinus maenas* (Kegel et al. 1989). Its identification also occurred in the lobster *Homarus americanus*, followed by the crayfish *Orconectes limosus* (Kegel et al. 1991; Chang, Prestwich, and Bruce 1990). The moult inhibiting hormone (MIH) was first isolated from *C. maenas*, the gonad inhibiting hormone (GIH) was identified in the species *H. americanus* and the mandibular organ inhibiting hormone (MOIH) was later found in the crab *Cancer pagurus* (Webster 1991; Soyeux et al. 1991; Wainwright et al. 1996). The neuropeptide hormones are mainly constituted by 72 to more than 80 sequenced amino acids, presenting a specific and invariant signature characterized by six well-conserved cysteine residues forming three intra-molecular disulfide bridges (Chen, Toullec, and Lee 2020). However, further characterisation of available sequences revealed major sequence similarity among MIH, GIH and MOIH hormones compared to CHH (Chen, Toullec, and Lee 2020), giving rise to two subfamilies in the CHH-superfamily: type-I comprising CHH/ITP, and type-II comprising MIH, MOIH, GIH (Webster, Keller, and Dirksen 2012). The ion transport peptide (ITP) is an ortholog of CHH identified in insects. The CHH-superfamily peptides are not confined to the XO-SG system and the pericardial organs. Their expression equally occurs in neural and non-neural tissues and organs such as *C. maenas* fore- and hindgut epithelium (Chung, Dirksen, and Webster 1999).

CHH type-I neuropeptide's defining role is related to controlling carbohydrate metabolism (Webster, Keller, and Dirksen 2012). When secreted, phosphorylase and hyperglycaemia are activated. In other words, glycogen synthase is inhibited.

Consequently, a mobilisation and concentration of glycogen in target tissues occur (Leuven et al. 1982; Sedlmeier 1985). The adaptive significance of hyperglycaemia is evident when CHH rapidly increases in crustaceans exposed to perceived stressful factors and pollutants (Chang, Keller, and Chang 1998). Crustacean hyperglycemic hormone modulates anaerobic energy metabolisms such as glycolysis, a process adapted by crustaceans facing stressful environmental or physiological conditions (Chung et al. 2010). In the shore crab *Carcinus maenas*, previous studies demonstrated that CHH can inhibit Y-organ production of ecdysteroids (Hopkins 2012), stimulate amylase release from the midgut gland and elevate free fatty acids and phospholipids (Santos et al. 1997; Sedlmeier 1988). The gonad inhibiting hormone or vitellogenesis inhibiting hormone (VIH) is directly involved in the endocrine control of the crustaceans' reproduction process, negatively regulating gonad maturation. In females, it inhibits vitellogenin (a yolk protein precursor) accumulation during oocyte development and ovarian maturation. At the same time, in males, it affects the secretion of an androgenic gland (AGH) responsible for the normal development of the male secondary sexual characters and reproductive system (Treerattrakool et al. 2008; Fingerman, Jackson, and Nagabhushanam 1998). In contrast, MOIH represses the synthesis of methyl farnesoate (MF) by the mandibular organ (MO), known as precursor of insect juvenile hormone III (JH III), involved in the morphogenesis of the male propodus and ovary growth in females, influencing early stages of vitellogenesis (Wainwright et al. 1996; Laufer, Biggers, and Ahl 1998). This indicates that mandibular organs rapidly activate in the absence of this inhibitory hormone, while MF synthesis is continuously modulated by its presence (Wainwright et al. 1996).

Decapod crustaceans' growth is thus under endocrine and hormonal control. The old hardened exoskeleton or cuticle is periodically shed and renewed, a cyclic process called moulting or ecdysis (Knigge, LeBlanc, and Ford 2021). The moult cycle comprises four major stages regulated by ecdysteroid hormones synthesised by the Y-organs: intermoult, proecdysis or premoult, ecdysis and post-moult (Spindler, Adelung, and Tchernigovtzeff 1974). Previous experiments involving eyestalk ablation confirmed the Y-organs as the moulting glands in the crab *Carcinus maenas* (Chang and Mykles 2011). Y-organs are negatively regulated by the neuropeptide CHH and the moult-inhibiting hormone (MIH). This interaction drives the progression through the moult cycle. During intermoult, pulsatile releases of MIH maintain Y-organs in a basal state with repression of ecdysteroid synthesis, preventing entry into premolt (Chung et al. 2010). Contrarily, a reduction in circulating MIH relieves Y-organ repression allowing the transition to the synthesis of the new exoskeleton outer layers and partial resorption

of the old cuticle (Chang and Mykles 2011). At the end of premolt, a rapid drop in haemolymph ecdysteroid levels triggers the release of the animal from its carapace, known as ecdysis (Knigge, LeBlanc, and Ford 2021; Mykles and Chang 2020). Ecdysone is one of the most important steroid moulting hormones regulating ecdysis, peripherally transformed into the active hormone 20-hydroxyecdysone (Mykles 2011).

1.1.4. Model organism in ecotoxicological studies

Carcinus maenas, one of the best-studied estuarine species, has been widely used as a model organism in aquatic ecotoxicology studies given its well-known biology, ecology and physiology, broad geographical distribution, easy capture and maintenance in the laboratory (Rodrigues and Pardal 2014). Additionally, its sexual dimorphism constitutes an advantage in evaluating specific responses related to gender (Pereira et al. 2011). This macroinvertebrate is described in multiple research works as a suitable sentinel to assess the quality of estuarine ecosystems, and it occupies a key position in marine food webs connecting primary producers and top predators (Rodrigues and Pardal 2014; Mesquita, Guilhermino, and Guimarães 2011). Furthermore, green crab communities are essential for evaluating benthic integrity, recommended for inclusion in monitoring programmes (Picado et al. 2007), and are important biological indicators of the impact of environmental contamination (Rodrigues and Pardal 2014; Bamber and Depledge 1997). Indeed, several studies demonstrated the value of *C. maenas* to assess the potential impact of various heavy metals (Cd, Cu, Co, Zn and others), pharmaceutical drugs, pesticides, polyaromatic hydrocarbons (PAHs) and polychlorinated biphenyls (PCBs) discharged along its natural habitats, from marine activities, agricultural areas and industrial and urban wastewaters (Rodrigues and Pardal 2014; Pedersen et al. 1994; Galloway et al. 2004). When exposed to even moderate pollution, biochemical alterations in *C. maenas* demonstrate sensitivity to abiotic and chemical stress, which is useful in biomarker approaches to characterise the resultant effects from exposure to single contaminants or their cocktails (Leignel et al. 2014; Rodrigues et al. 2015). The measurement of various parameters has proven helpful to these evaluations, such as cardiac rate, cellular viability and stability, gene expression, neurotoxicity and oxidative stress responses, locomotory and feeding behaviour (Mesquita, Guilhermino, and Guimarães 2011). The mentioned characteristics make *C. maenas* a suitable crustacean species to investigate the effects of environmental contaminants and a good biological indicator for assessing the health status of estuarine and coastal ecosystems.

1.2. Estuaries of Minho and Lima Rivers

Estuarine ecosystems are transitional zones between marine and freshwater domains (Sousa et al. 2008), characterized by distinct ecological and environmental functions (e.g. subsistence of biodiversity, mitigation of floods and movement of nutrients, carbon sequestration and storage) and high vulnerability to sea level rise. Estuaries are particularly exposed to coastline erosion and natural phenomena such as storms and floods (Pinto and Martins 2013; Martins et al. 2019; Meire et al. 2005). They also constitute essential reproductive and vital nursery grounds for several species, including crustaceans. Estuarine ecosystems are areas of remarkable biological productivity with high conservational significance (Meire et al. 2005). They are also preferred for urban, industrial and agricultural settlement and development, given its economic and social value (Sousa et al. 2008; Guimarães, Medina, and Guilhermino 2012). This persistent anthropogenic pressure along coastal waters promotes an excessive daily production and discharge of chemical contaminants, of which a considerable part accumulates in sediments, impacting the survival of the inhabiting organisms, local biodiversity and the ecosystems' functioning (Gravato et al. 2010; Rodrigues et al. 2012).

The Continental Portugal (Iberian Peninsula, SW-Europe) coastline is vast, extending over 900 km and presenting heterogeneous geomorphological features that include, among others, rocky coasts, sandy shores backed by dunes, beaches and bays, lagoons and estuaries (Pinto and Martins 2013). In the western Iberian margin, the Minho river is located on the border between the Minho (north of Portugal) and Galicia (Spain) regions (Mil-Homens et al. 2013b). It rises at an altitude of 750m in Serra de Meira (Galicia) and meets the Atlantic Ocean between the A Guarda (Galicia) and Caminha (Portugal) region (Mil-Homens et al. 2013a). It is 300km long, with its mesotidal estuary constituting approximately the last 40km of the river, as far as the influence of tides extends (Sousa, Guilhermino, and Antunes 2005). The total estuarine area is 23km², with the seawater and freshwater mixing in the lower portion. The estuarine basin is characterized by a tighten channel with significant water flow, presenting high hydrodynamic energy and lower water residence time (Saraiva et al. 2007).

The Minho river estuary is a protected area with wetland habitats of high ecological importance (Pinto and Martins 2013). Its predominant Atlantic climate with some Mediterranean influence creates and maintains favourable conditions to the occurrence of bird and macroinvertebrate communities and fish populations, including migratory

fish species (Pinto and Martins 2013; Dias et al. 2012; Mota et al. 2015; Souza et al. 2013). Due to its relatively low human intervention, it is included in protection networks that seek to secure habitats and biodiversity from threats and impacts, including those related to climate change (Sousa et al. 2008; Guimarães, Medina, and Guilhermino 2012; Sousa, Guilhermino, and Antunes 2005). Since it is considered as low impacted with reduced contamination levels, Minho river estuary is used as a reference in ecotoxicological studies despite not being absent a few localized sources of pollution, as areas of forestry and irrigated agriculture, population settlement at its basin, nautical activities and relatively low industrialization (Pinto and Martins 2013; Ferreira 2003; Guimaraes et al. 2009; Monteiro et al. 2007).

The Lima river is located south of the Minho river, draining near the city of Viana do Castelo (Portugal) into the Atlantic ocean (Ramos et al. 2006). It rises in the province of Ourense (Galicia, Spain) at an altitude of 975m and has an extent of 108km, of which 67km belong to the Portuguese territory. Its mesotidal estuary, located in the Northwest of Portugal, is 20km long and comprises a total estuarine area of approximately 6km² (Costa-Dias, Sousa, and Antunes 2010). This river is included in the CORINE biotopes project (Moss and Wyatt 1994), responsible for gathering and ensuring information consistency on the state of its environment and natural resources, since it comprises important and sensitive areas (e.g. salt marsh, sand flats and mud flats), valuable for species and habitats conservation along the estuary (Costa-Dias, Sousa, and Antunes 2010). Several species communities occur in this river, enhancing marine fish, shorebirds, mammals and macroinvertebrates (Antunes and Dias 2005; Sousa 2003). Compared to the Minho estuary, the Lima estuary is considerably impacted by anthropogenic activities, with pressures occurring from a paper mill located upstream, a marina and particularly from a harbour and a shipyard established near the mouth (Guimarães, Medina, and Guilhermino 2012; Cairrao et al. 2004). In addition to these infrastructures, nutrients and several pollutants resulting from urban and industrial sewage discharges and agricultural runoff are transported into the estuarine area (Azevedo et al. 2013). Previous research reported the presence of polycyclic aromatic hydrocarbons (PAHs) and heavy metals collected from its sediments, impacting water quality and the health status of the local organisms (Guimaraes et al. 2009). Additionally, prior studies have classified the Lima estuary as moderately disturbed, given the predominance of opportunistic species and the noticeable environmental stress (Costa-Dias, Sousa, and Antunes 2010; Sousa, Dias, and Antunes 2007). Further investigation on potential differential sensitivity to contaminants exposure in a

moderate contamination level estuary and pollutants effects, particularly in key structuring species, such as the green crab *Carcinus maenas*, is still required.

Figures 4 and 5 illustrate the geographic location of both Minho and Lima rivers, and the Lima estuary disposition, respectively.

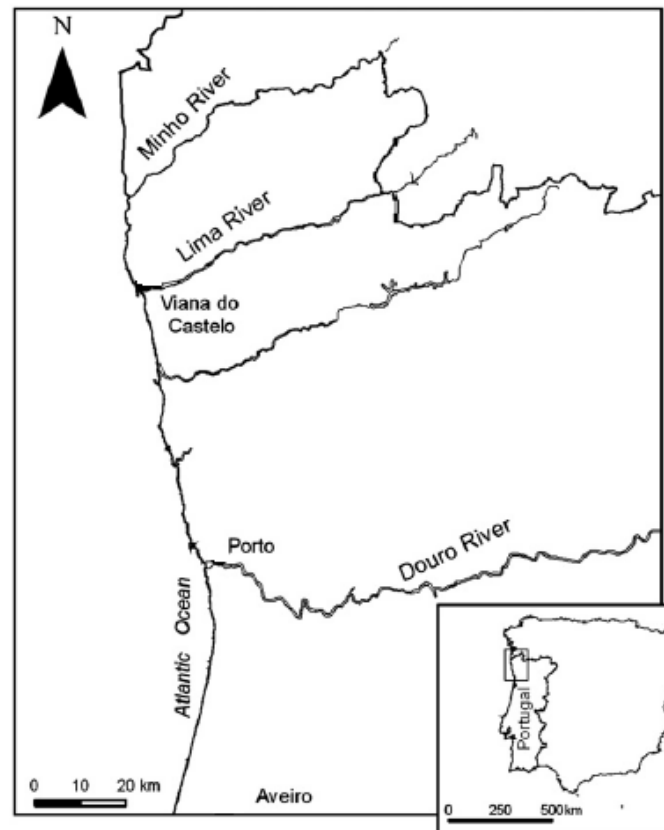


Figure 4 – Localization of Minho and Lima rivers in the Northwest coast of Portugal. Image retrieved from (Gravato et al. 2010).

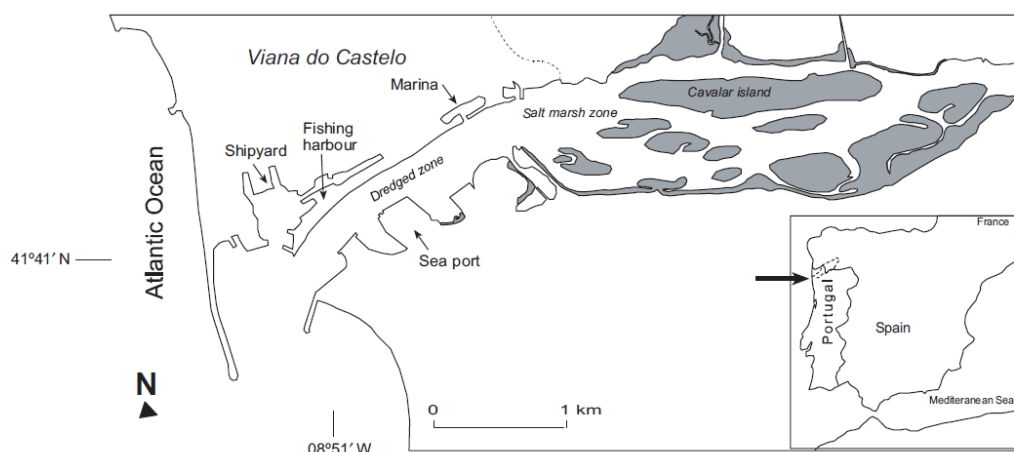


Figure 5 – Lima river estuary. Disposition of the anthropogenic structures developed along the estuary. Sand islands and salt marshes represented as grey areas. Image retrieved from (Ramos et al. 2006).

Previous studies have been demonstrating that the green crab populations of these estuaries show differential sensitivity to chemical contamination and abiotic stress, suggesting that the exposure history in Minho and Lima estuaries may influence their response to contamination (Rodrigues, Santos, et al. 2014; Rodrigues, Gravato, and Guimaraes 2013; Rodrigues et al. 2015; Rodrigues et al. 2012). Furthermore, the results of those studies suggest such differential responses occur via physiological acclimation, rather than adaptation.

2. Sequencing

2.1. Initial advances in sequencing technologies

Over the past decades, genome and transcriptome studies have been revolutionized by the considerable progress of sequencing methods. Low-throughput techniques, such as northern blots and quantitative polymerase chain reaction, were firstly used in gene expression studies (Kukurba and Montgomery 2015). Hybridization-based microarray technologies posteriorly succeed, allowing the quantification of multiple RNA molecules, and the simultaneous measurement of the expression levels of thousands of genes, at relatively lower costs (Schena et al. 1995). Microarrays are still widely used in genomic and transcriptomic research. However, some associated limitations hamper its application in this field, for instance: the need for a priori knowledge of the sequences under analysis, limited accuracy in quantifying little and highly expressed genes, and problematic cross-hybridization artefacts when studying

very similar sequences (Casneuf, Van de Peer, and Huber 2007; Shendure 2008). To overcome these issues, sequence-based approaches started to be developed and introduced. The Sanger sequencing strategy was the ‘first-generation’ technology to directly elucidate transcriptomes by determining the transcript sequence (Metzker 2010). In a simplified way, this method amplifies the sequence to be determined into millions of copies and synthesizes complementary templates using known primer sequences and a combination of deoxy-nucleotides (dNTPs), and dideoxy-nucleotides or ddNTPs (Sanger, Nicklen, and Coulson 1977). This mixture between deoxy- and dideoxy-nucleotides randomly and non-reversibly terminates the chain extension process, creating extended molecules with different lengths (Kircher and Kelso 2010). These fragments are then partitioned and analysed to reveal the final sequence. Primarily, this separation was achieved according to their weight through gel electrophoresis (Swerdlow and Gesteland 1990; Kircher and Kelso 2010). However, developments in fluorescent dyes and fluorescence detection, evolved this strategy to the known ‘automated Sanger sequencing’ (Metzker 2005). In this case, fluorescent dyes are used to label either the primers or the terminating ddNTPs. These tagged molecules are excited when passing through a laser-detection region, producing fluorescence emissions of four distinct colours (Metzker 2005). Base pairs are then orderly aligned through their associated colouration, revealing the ultimate sequence (Metzker 2005; Kircher and Kelso 2010). This conventional approach, presented in 1977, dominated the industry for many years, having been the first to complete the human genome sequence (Consortium 2004). This accomplishment allowed and encouraged the sequencing of the genomes of several model organisms, e.g. mouse and chimpanzee (Sequencing and edu 2005; Simon 3 et al. 2002). However, few advances have been reported about the Sanger sequencing strategy since it presents some disadvantages compared to more recent technologies. In the case of transcriptomes, its limited-throughput affects the quantification of transcripts, constituting an evident setback in the sequencing field (Reuter, Spacek, and Snyder 2015). Moreover, its higher-cost, incapacity to measure expression levels of splice isoforms, and requirements for large amounts of input RNA have been restricting the use of this approach (Reuter, Spacek, and Snyder 2015; Kukurba and Montgomery 2015).

2.2. High-throughput Next Generation Sequencing methods

First-generation methods began to be surpassed with the introduction of high-throughput next-generation sequencing (NGS) technologies. The development of these

newer strategies revolutionized the transcriptomics field by allowing the analysis of RNA through the sequencing of complementary DNA (cDNA) at a higher scale, also known as RNA-Seq (Ozsolak and Milos 2011). A major improvement provided by NGS was the potential to generate massive volumes of data inexpensively (Metzker 2010). Through a parallelized process, high-throughput technologies may sometimes exceed one billion short-reads per run (Metzker 2010). Though these reads endured a substantial reduction in their lengths, assembly processes (e.g. *de novo* assembly) in such data are still very accurate, given the provided high-coverage sequencing (Hutchison III 2007). This expanded the use of these approaches from large to small laboratories, replacing Sanger sequencing and microarrays methodology in multiple research. For instance, gene-expression studies were improved with the possibility to identify and quantify transcripts without prior knowledge of a specific gene (Harrison, Wright, and Mank 2012). Determine events of variation, allele-specific expression, and alternative splicing in identified genes became equally possible (Kukurba and Montgomery 2015). In comparative and evolutionary studies, it became achievable to sequence the whole genome of many related species (Hudson 2008). Understanding the nature of transcriptomes and aspects of transcript biology was enhanced through RNA sequencing (RNA-Seq), in both prokaryotic and eukaryotic organisms (Metzker 2010; Wang, Gerstein, and Snyder 2009). Deeply profiling their transcriptomes allowed to clarify and associate physiological and pathological conditions (Kukurba and Montgomery 2015). Besides transcriptomics, NGS technologies also had a substantial impact on barcoding and environmental monitoring (Carew et al. 2013). They allowed the retrieve of sequence data directly from environmental samples, as well as sequencing multiple organisms from diverse samples simultaneously (Pfrender et al. 2010; Hajibabaei et al. 2011). This was particularly considered for taxonomic groups (e.g. invertebrates) occupying estuarine or marine habitats to assess the condition of these ecosystems accurately and determine the influence of pollutants at a transcriptomic level (Hajibabaei et al. 2011; Chariton et al. 2010). As with any other strategies, a few drawbacks are associated with NGS technologies. For instance, data have associated higher error rates (Liu et al. 2012). Also, these require innovative storage and bioinformatic solutions at a rapid pace, for example, in mapping short-reads against a reference sequence or *de novo* assembly (Pop and Salzberg 2008). Reads with smaller lengths require careful examination, especially for variant discovery (Goodwin, McPherson, and McCombie 2016).

2.3. High-throughput Next-Generation Sequencing platforms: second and third generations

The first available NGS platform, GS 20, was developed by 454 Life sciences based on a 'pyrosequencing' technique (Margulies et al. 2005; Nyrén, Pettersson, and Uhlén 1993; Ronaghi et al. 1996). The Genome Sequencer instrument allowed shotgun sequencing to be performed without cloning any host cell (Hutchison III 2007). In the following years, Roche applied science acquired 454 Life sciences, extending the GS instrument to a new version known as GS FLX titanium (Pareek, Smoczynski, and Tretyn 2011). This expansion and improvement introduced the second generation of high-throughput next-generation sequencing (HT-NGS) platforms. Besides Roche, generating approximately five hundred million bases of raw sequence, Illumina and SOLiD technologies were developed and presented, producing billions of bases in a single sequencing run (Pareek, Smoczynski, and Tretyn 2011). These novel methods stood out by substituting the 'pyrosequencing' technique with a parallelized process, involving clonal amplification (Reuter, Spacek, and Snyder 2015). SOLiD (*Supported Oligonucleotide Ligation and Detection system*) focused on hybridization-ligation chemistry, using sequential ligation of oligonucleotides probes to sequence (Shendure et al. 2005; Pareek, Smoczynski, and Tretyn 2011). On the other hand, Illumina relied on chain-terminating nucleotides and reversible dye terminators (Metzker 2010). Though each second-generation platform presents particular specificities, its methodological steps are equivalent: 1. preparation of a template (DNA, cDNA); 2. massive parallel process using PCR for clonal amplification; 3. sequencing and alignment of short-reads. Figure 6 illustrates the main second-generation platforms leading the sequencing field.

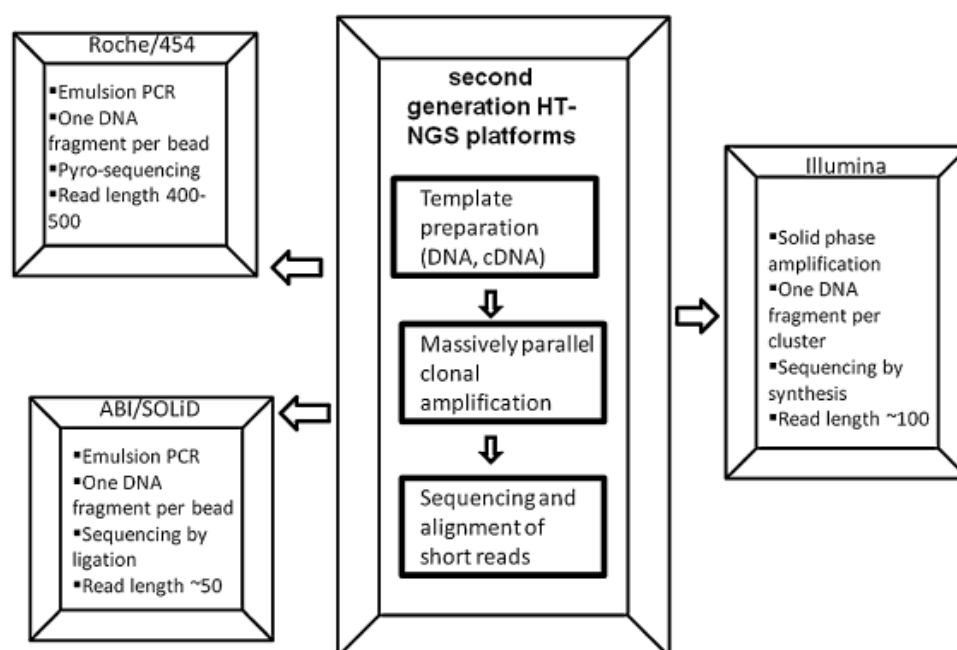


Figure 6 – The second-generation HT-NGS platforms leading the sequencing field. Image retrieved from (Pareek, Smoczynski, and Tretyn 2011).

A third-generation of HT-NGS platforms began to appear rapidly, to outline a few biases introduced during the PCR amplification phase, for instance: base sequence errors or distortion of abundance and frequency at a sequence level (Pareek, Smoczynski, and Tretyn 2011). The novelty of these methods was the introduction of the sequencing-by-synthesis concept, removing the initial amplification step (Ansorge 2009). In this case, a sequence is determined directly through a single DNA molecule. Pacific Biosciences introduced the first platform using this new concept, known as Single Molecule Real-Time (SMRT) sequencer (McCarthy 2010). Following, other platforms and technologies began to appear, leading the field of third-generation sequencings, such as 1. Heliscope sequencer introduced by Helicos biosciences; 2. Nanopore sequencer from Oxford Nanopore; 3. Real-Time Single Molecule sequencer developed by VisiGen Biotechnologies; 4. Ion Torrent technology a product of Thermo Fisher Scientific (Hong et al. 2020). Third-generation platforms present particular specificities with methodological equivalent steps involving: 1. preparation of a template (DNA, cDNA); 2. incorporation of nucleotides with fluorescent labelling; 3. real-time sequencing of a single molecule. Figure 7 illustrates the main third-generation platforms leading the sequencing field.

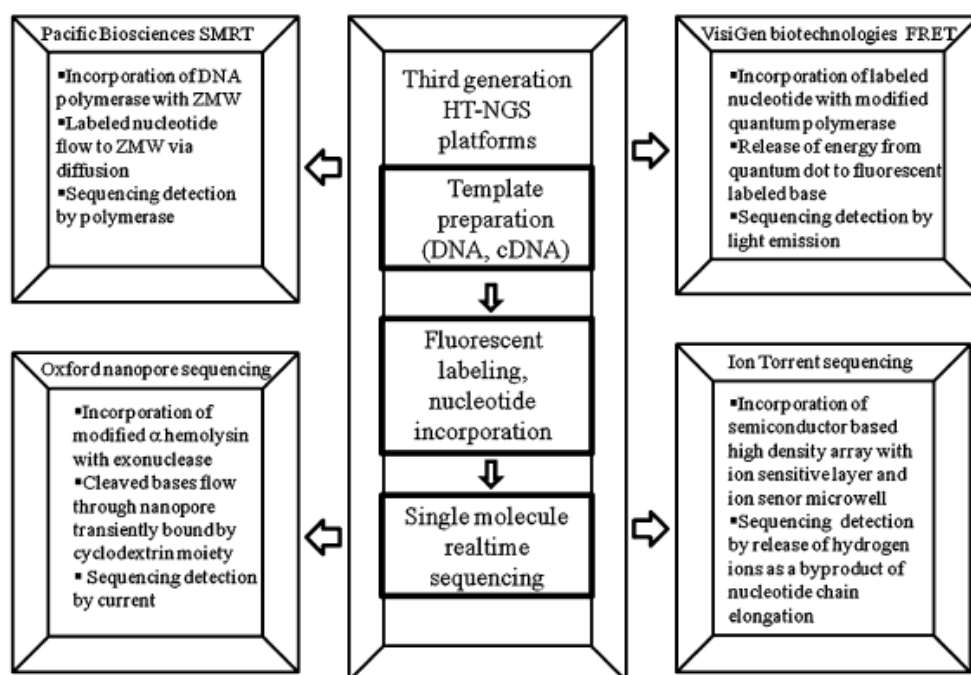


Figure 7 – The third-generation HT-NGS platforms leading the sequencing field. Image retrieved from (Pareek, Smoczynski, and Tretyn 2011).

Common to both generations is the preparation of a template, involving isolation and extraction of biomolecules (e.g. DNA, RNA or protein), followed by the creation of a sequencing library. In RNA-Seq experiments, the extracted molecule is RNA, posteriorly converted into cDNA to construct the library. Initially, RNA is isolated from a biological sample for transcriptome sequencing. This unstable molecule is usually extracted with the utilization of 4M guanidinium thiocyanate or phenol and SDS solutions (Doyle 1996). As an example, the TRIzol reagent combines phenol and guanidinium isothiocyanate to purify RNA. As described in (Rio et al. 2010), this monophasic solution denatures proteins and solubilizes biological material from a single sample. Once extracted, RNA quality is usually determined through an RNA integrity number (RIN), ranging from 1 to 10, provided by an Agilent Bioanalyzer (Kukurba and Montgomery 2015). Higher quality values (>6) indicate the least degradation, avoiding erroneous biological conclusions in further analysis (Kukurba and Montgomery 2015). As mentioned above, a biological sample is usually the source to extract RNA molecules. Samples-source selection is not simple, since an organism presents hundreds of cell types, tissues, and differential developmental stages with unique gene expression signatures (Kukurba and Montgomery 2015). Ultimately, this choice commonly depends on the research objectives and samples' availability.

Another aspect that must be considered in template preparation is the heterogeneity of the samples since low purity levels interfere with the accuracy of gene expression quantification (Kukurba and Montgomery 2015). Various purification protocols and kits are available, such as the 'RNeasy Mini Kit®'. Afterwards, a sequencing library is constructed, requiring the selection of a particular RNA species. There are multiple types of naturally occurring RNA, including small RNAs, transfer RNA (tRNA), messenger RNA (mRNA) and ribosomal RNA (rRNA), with the predominance of the latter in the total RNA pool (Buckingham 2019). Before creating a library, rRNA transcripts are usually removed since they may prevent the detection of other less-abundant RNAs (Kukurba and Montgomery 2015). Once RNA molecules are selected and isolated, its conversion into cDNA takes place. Briefly, the first complementary DNA strand is synthesized from an RNA template, via reverse transcription, using random or oligo(dT) primers (Stark, Grzelak, and Hadfield 2019). The second cDNA strand is generated by using the first one as the reference. Double stranded cDNA molecules are then fragmented, and adapters are ligated to their extremities for amplification (Stark, Grzelak, and Hadfield 2019). Lastly, a sequencing library, deriving from each cDNA strand and thus representing the parent mRNA strand and its complement, is constructed (Kukurba and Montgomery 2015).

2.4. Illumina Sequencer

As mentioned above, the Illumina sequencer, formerly known as Solexa, belongs to the second-generation of HT-NGS technologies, introducing its first platform (Genome Analyzer) in 2007 (Shokralla et al. 2012). Its workflow differs by implementing a sequencing-by-synthesis approach coupled with bridge amplification on the surface of a flow cell (Shokralla et al. 2012). Initially, DNA molecules are partitioned, and adaptors are ligated to both ends of the generated fragments (Stark, Grzelak, and Hadfield 2019). This allows the construction of an Illumina-specific adapter library. Single-stranded library fragments, obtained with the application of sodium hydroxide, posteriorly connect to the eight channels constituting each flowcell through active heating and cooling steps (Shokralla et al. 2012; Kircher and Kelso 2010). These channels (or lanes) are composed of two populations of covalently attached oligonucleotides, complementary to both adapter ends of the sequencing library (Kircher and Kelso 2010). Multiple binding combinations can be performed. For instance: a singular library may be ligated to all eight channels, or different libraries can be added to each of the eight lanes (Mardis 2008). Once DNA fragments and

oligonucleotides hybridize, a bridge amplification process, using long primer sequences, occurs repeatedly. Following, randomly scattered clusters containing millions of copies of the original fragments are formed (Kircher, Heyn, and Kelso 2011). This process is achieved using an automated device called Cluster Station (Mardis 2008). Figure 8 summarizes the sequencing-by-synthesis approach used by the Illumina sequencer.

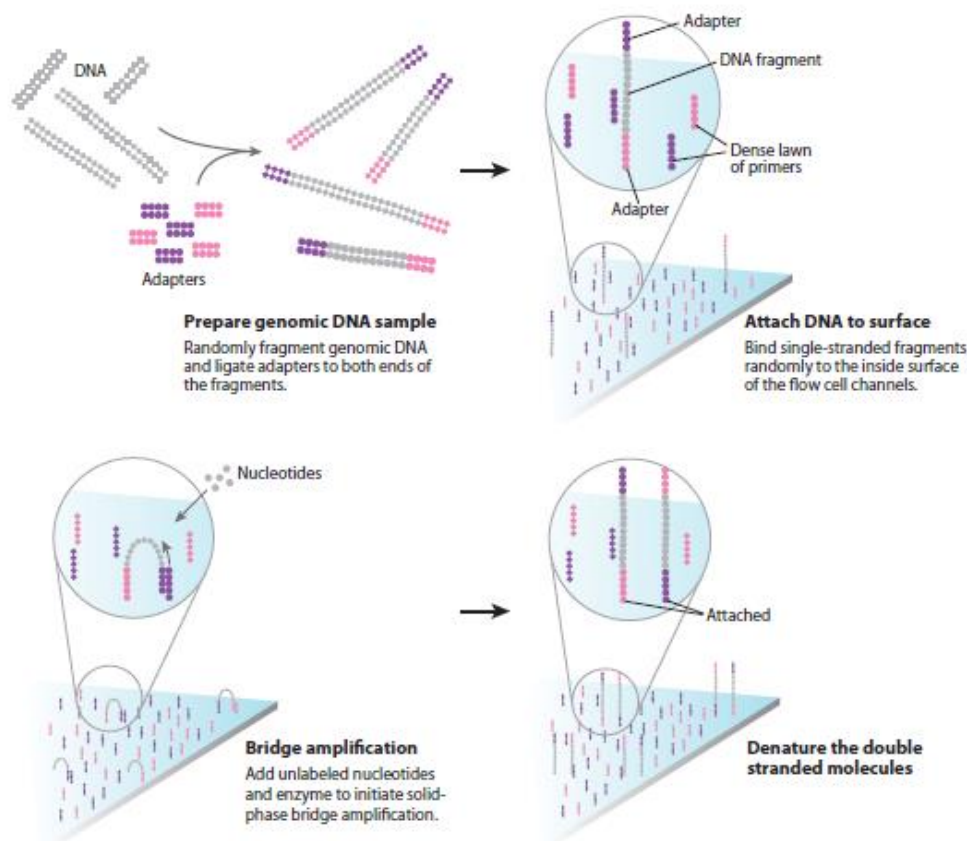


Figure 8 – Sequencing-by-synthesis approach used by the Illumina sequencer. DNA partition and ligation of adapters to fragments extremities. Hybridization among fragments and oligonucleotides, in a flowcell. Lastly, bridge amplification. Image retrieved from (Mardis 2008).

Each cluster is supplied with a DNA polymerase and four nucleotides differentially labelled with unique fluorescents (Shokralla et al. 2012). Similarly to the Sanger sequencing strategy, each nucleotide is incorporated by the polymerase during the sequencing primers extension (Fuller et al. 2009). Since its 3'-OH extremities are blocked and inactivated, only a single nucleotide base is added per flow cycle (Shokralla et al. 2012). This cyclic reversible termination strategy is then followed by

laser excitation and imaging processes, which are applied to capture and read each nucleotides fluorescent-label signal (Bentley et al. 2008). Once a cycle is concluded, the 3' blocking groups are removed, allowing the incorporation of the following nucleotide in the succeeding flow cycle (Mardis 2008; Kircher, Heyn, and Kelso 2011).

Illumina has a suite of sequencers available optimized for distinct through-puts and turnaround times (Reuter, Spacek, and Snyder 2015). Besides Genome Analyzer, HiSeq and MiSeq are well-established platforms, producing vast amounts of output data (superior to 300Gb), in a quite reduced time (Reuter, Spacek, and Snyder 2015). The series of HiSeq platforms has been constantly evolved and upgraded, providing multiple versions and features for sequencing, for instance: the HiSeq 1000, HiSeq 2000 and HiSeq 2005. Today, the latter generates up to 120 Gb of data in little more than a day, at a lower cost, enabling an entire genome to be sequenced in 24h (Shokralla et al. 2012). HiSeq 3000 and HiSeq 4000 are already available. MiSeq is also a high-capacity sequencer, generating up to 300bp reads, with a total throughput of 1.5–2 Gb per run (Shokralla et al. 2012). Given the speed of its run times (as low as 4 hours), the generated outputs are commonly intended for small genomes and targeted sequencing (Reuter, Spacek, and Snyder 2015). Among those, NextSeq 500 and HiSeq X are the most novel strategies presented by Illumina. NextSeq is similar to the MiSeq platform. However, it produces 120Gb in less than 30 hours (Reuter, Spacek, and Snyder 2015). Moreover, it substitutes the four-channel sequencing method with a two-channel strategy. In this case, only two images are required for nucleotide detection, reducing data processing times (Reuter, Spacek, and Snyder 2015). The HiSeq X exclusively supports whole-genome sequencing of human samples, producing 1.8 Tb of data in 3 days. Additionally, its incorporation of a new pattern flow cell technology allows for an enormous rise of throughput and superior cluster densities (Reuter, Spacek, and Snyder 2015). As with other strategies, a few errors are associated with Illumina platforms. Though its accuracy rate exceeds 99%, under-representations are still displayed in AT and GC regions, mostly due to amplification bias during template preparation (Dohm et al. 2008; Hillier et al. 2008; Wall et al. 2014). Substitution errors are also very commonly found (Harismendy et al. 2009). It is also evidently known that sequencing errors increase with each runned cycle. However, Illumina techniques allow to sustain up to 150 sequencing cycles with reasonable error profiles, with a mean error and last cycles errors below 1% and 10%, respectively (Kircher, Heyn, and Kelso 2011).

Illumina sequencing platforms have been highly efficient in several RNA-Seq studies. For instance, its generated short-reads are considered suitable for gene

expression profiling (Pareek, Smoczynski, and Tretyn 2011). Research for annotation, the discovery of small RNA molecules from transcriptomic data, and the detection of gene fusions are also favourable with the application of Illumina sequencers (Yang et al. 2010). Another advantage it is the inclusion of reverse strand sequencing, also known as the paired-end sequencing approach. This strategy allows Illumina platforms to produce sequence data in duplicate amounts (Kircher, Heyn, and Kelso 2011). Moreover, mapping and assembly processes may be improved when the distance between the obtained pair-ends is known (Chaisson, Brinza, and Pevzner 2009).

2.5. High-throughput sequencing quality assessment

High-throughput sequencing (HTS) technologies have been introduced at an unprecedented pace in genome and transcriptome studies, with sequence data being massively generated in a single experimental run (Metzker 2010). Its cost-effectiveness and improved computational speed allowed sequencing to become a routine application in small to large laboratories or institutes (Lo and Chain 2014). RNA sequencing relies on these technologies for gene and allele specific expression studies and alternative splicing analysis, often replacing or superseding microarrays and exon arrays, respectively (Wang et al. 2008; Wang, Gerstein, and Snyder 2009; Montgomery et al. 2010). Likewise, it commonly uses its tools for more profound and precise identification of novel transcripts, unknown transcribed regions and gene fusion events (Denoeud et al. 2008; Maher et al. 2009).

However, a considerable amount of errors arise frequently during library constructions and sequencing processes, negatively impacting the quality of the raw data for downstream analysis (Trivedi et al. 2014). In RNA-Seq data, the presence of transcript artefacts is quite noticeable, with each HTS platform displaying a corresponding unique error pattern (Dohm et al. 2008). Regardless of the sequencing platform, these sequence artefacts are generally identified based on its origin of appearance:

1. Sequencing process: as a run progresses, sequence reads quality starts degrading. For instance, RNA degradation severely impacts the accuracy of subsequent transcripts analysis results. Errors may also occur along the sequences and at their bases, due to: insertions/deletions, low-quality base calls, and variable quality scores among the sequence reads (Patel and Jain 2012; Zhou et al. 2018);

2. Library construction: e.g.: contamination of inserted technical sequences (primers/adapters), duplicates introduced by PCR amplification bias, irregular read coverage (Zhou et al. 2013);
3. Outside sources: including known and unknown species. For instance: contamination during sample preparation, nucleotides from unexpected species already present in the samples (Zhou et al. 2013).

Ignoring sources of contamination among the data may mislead biological interpretations and conclusions, impacting results reliability. Several computational methods and software tools have been developed to prevent and outline these limitations to detect and pre-process sources of low-quality and contamination among the raw data. As a first step, quality control (QC) assessments were implemented to identify the intrinsic quality of the raw sequence reads through its base composition or assigned quality scores (Trivedi et al. 2014). The FastQC tool is highly used as an initial QC checkpoint to evaluate the overall quality of a sequencing run (Trivedi et al. 2014). Afterwards, most tools and pipelines implement read trimming on the initial raw data. Trimming is a pre-process strategy that estimates error rates from a read, eliminates only lower-quality portions, and maintains longest higher-quality regions, preserving biological information (Del Fabbro et al. 2013). Examples of positive effects resultant from the application of read trimming steps are demonstrated in a research work developed by (Del Fabbro et al. 2013). It highlights the beneficial influence of trimming prior to assembly approaches, genotyping analysis (e.g. SNP identification) and RNA mapping when comparing multiple algorithms with trimming functions. For example, it demonstrates how the no removal of low-quality reads during a genome or transcriptome assembly induces the appearance of false k -mers (substrings of length k comprehended within a biological sequence), increasing the complexity of its following processes (Chikhi and Medvedev 2013). Further, it exhibits how executing trimming steps lead to a reduction in the number of initial raw reads, increasing the remaining sequences' capability to align over a reference genome or transcriptome correctly.

Several applications provide trimming features, data filtering and quality assessment steps. Trimmomatic is a tool optimized to pre-process HTS data, with the capability to remove adapters previously inserted for the sequencing process (Bolger, Lohse, and Usadel 2014).

Currently, research studies in bioinformatic field, addressing sequencing errors, have been focused in developing new pre-processing tools, not only able in profile the quality of a sequencing run and implement read filtering and trimming, but also correct possible discovered errors (Chen et al. 2014). For instance, AfterQC tool introduced in

(Chen et al. 2017). Moreover, it has been essential to pursue software with superior versatile functionality and better run-time efficiency, as HTQC and NGS QC toolkits presented in (Yang et al. 2013) and (Patel and Jain 2012), respectively. Additionally, various research works mainly focus on designing programs specific to RNA-Seq data. Even though some already exist, there are still restrictions in sequence trimming, contamination removal and processing speed, thus the requirement for RNA pre-processing tools progress and improve. For example, (Zhou et al. 2018) developed the RNA-QC-Chain tool, and (Goncalves et al. 2011) presented an ArrayExpressHTS pipeline to tackle distinct aspects of RNA-seq data analysis and solve and outline these issues.

3. Thesis objectives

The main objective of the present thesis was to investigate and identify possible differential responses and affected metabolic pathways in the shore crab *Carcinus maenas* from estuarine environments with distinct levels of contamination through the bioinformatic analyses of high-throughput transcriptomic data. Transcriptomic responses of any organism, although complex, are one of the most accurate forms to determine the impact of stressors at a lower organizational level. Moreover, they can be helpful in distinguishing between adaption and acclimation (or physiological adaptation) phenomena. When organisms are exposed to an unknown or very complex mixture of contaminants, transcriptomes can allow to understand which pathways may be most impacted. Though *Carcinus maenas* populations from both studied estuaries seem well adapted to their living environment, previous data has shown that crabs respond differently to chemical and abiotic stress. The results of former studies also suggest that such responses should be due to acclimation elicited by the continuous exposure of crabs to chemical contamination in the Lima estuary compared to the low-impacted Minho estuary. This investigation was thus carried out to analyse high-throughput sequencing data obtained for the two populations and possibly clarify potential mechanisms involved in such differential responses.

The data was obtained from crabs previously collected from the north-western Portuguese Lima and Minho estuaries under different anthropogenic pressure: Lima river under moderate contamination and Minho river under lower levels of contamination and human intervention (Rodrigues, Oliva-Teles, et al. 2014). RNA extraction from *C. maenas* individuals from the two selected estuaries, and posterior sequencing following Illumina TrueSeq protocol, were conducted in previous research in METOX (CIIMAR), within the scope of the project CRABTHEMES (PTDC/MAR/71143/2006). Thus, the already available mRNA sequences were used for the present study with the following established methodology and computational steps:

- i. perform a quality control assessment on the sequence data using FastQC tool;
- ii. remove low score/poor quality reads and adapters previously inserted for the molecular sequencing using Trimmomatic;
- iii. conduct *de novo* Transcriptome assemblies using Trinity assembler;
- iv. assess the redundancy of the reconstructed transcriptome assemblies implementing EvidentialGene (or Evigene) software package;
- v. evaluate the completeness of the assembled transcriptomes using BUSCO tool;

- vi. obtain valuation metrics to assess the quality of each performed step mentioned above.

The next chapters present the methodology employed in more detail, the results obtained in this bioinformatic research and their discussion. Figure 9 is a succinct representation of the application of computational and bioinformatic tools in analysing and interpreting biological data.

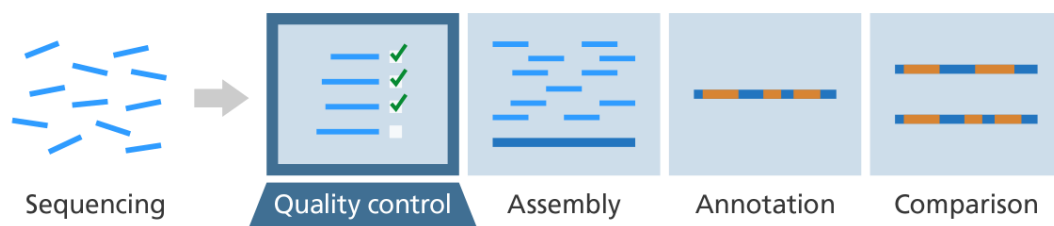


Figure 9 – Application of computational tools for analysis and interpretation of biological data. Image retrieve from the <https://www.yourgenome.org/> website.

Chapter II. Materials and Methods

1. Workflow I

1.1. Crab collection and estuaries sampling sites

The crab sampling campaign was carried out by (Rodrigues, Oliva–Teles, et al. 2014) in the Minho and Lima estuaries in August of 2012. *Carcinus maenas* individuals with complete appendices were collected at the mouth of each estuary during high tide, using trawl nets. Intermoult male crabs with similar colour morph, size and moult stage were selected for analysis and immediately transported to the laboratory. Once in the laboratory, crabs were allowed to deplete gut contents for 24h. After this period, the organisms were anaesthetised, measured and euthanised. Samples of the digestive gland were collected into RNeasy lysis buffer and kept at -80°C until RNA extraction and sequencing. Chemical analysis in sediments of these locations were previously described in (Guimarães, Medina, and Guilhermino 2012) for metals and polycyclic aromatic hydrocarbons, and in (Rodrigues, Oliva–Teles, et al. 2014) for metals. Metal content in tissues of *C.maenas* were also analysed in the study of (Rodrigues, Oliva–Teles, et al. 2014)

1.2. RNA extraction

RNA extraction from *C.maenas* individuals followed the methodology described in (Ferreira et al. 2019). Briefly, RNA was isolated and extracted using TRIzol® (Ambion) reagent, proceeded by a column purification step using RNeasy Mini Kit® (Qiagen - performed according to the manufacturer's manual) with the total RNA being stored at -80°C. Prior to freezing, RNA concentration and purity were assessed using a Nanodrop 2000c spectrophotometer (Nanodrop Technologies) and RNA integrity was verified using Bioanalyzer (RNA Nano Chip). A total of eight RNA sequence samples were extracted. Four samples (C3, C10, C13, G1R1) were collected from crabs of the contaminated Lima estuary, and four samples (R3, R11, R12, R15) were collected from crabs of the reference Minho estuary.

1.3. Library construction and Sequencing

Carcinus maenas RNA-seq libraries were prepared following Illumina TrueSeq protocol with an insert length of 150bp, and sequencing was performed on an Illumina HiSeq 2000. These processes were performed by Macrogen (Macrogen Europe BV, Netherlands). Paired-end sequences (20M reads/sample) were produced to derive a reference transcriptome.

2. Workflow II

2.1. Sequencing quality control using FastQC

A preliminary analysis of the initial generated raw sequence data was conducted since sequencing runs can display errors as a result of the biological processes of both library creation and the specifics of the sequencing reactions. Sequence quality control was analysed using the FastQC tool. Since sequences were aligned as read pairs, two input FastQ files, with forward and reverse reads, were analysed for each collected sample.

The FastQC software (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) performs quality control assessments on sequence data from high throughput sequencing pipelines using nine different modules (FastQC 2020). It provides a QC report, indicating possible problems or bias originated during the library construction or the sequencing process, followed by summary graphs and tables to visualize and interpret the data easily. Its analysis is usually fast since it uses a subsampling strategy in the total sequence dataset and exhibits the capability to place the generated results within a web server (Brown, Pirrung, and McCue 2017; Lo and Chain 2014). This tool, implemented using java language, accepts files in various formats, such as FastQ, SAM and BAM, GZip compressed FastQ, and others (Yang et al. 2013).

As mentioned, FastQC software uses nine different composition statistics modules to infer the sequence data condition and quality. The modules can be separated into two categories given the analysis target: analysis per base or full sequence analysis. Per base modules include:

1. Per base sequence quality - overview of all bases at each position in the input file (commonly, FastQ format) and its range of quality scores;
2. Per base sequence content – indicates the percentage of DNA bases (A, T, G and C) per base position;
3. Per base GC content – overview of the percentage of GC at each base position;

4. Per base N content – an N appears in a sequence when there was insufficient confidence during sequencing to make a base call. This module indicates the percentage of Ns per base position.

When considering the full sequence as the target of analysis, the following modules and its functions and features are considered:

1. Per sequence quality scores – determine if a subset of sequences has low quality scores;
2. Per sequence GC content - overview of the percentage of GC across each sequence total length;
3. Sequence length distribution – overview of the distribution of fragments with different sizes along all sequences;
4. Duplicate sequences – counting of duplication level per each sequence in a set (elevated duplicates may be indicative of bias);
5. Overrepresented sequences – listing of sequences appearing more than 0.1% in the total set (may be indicative of contamination). To this module, a k-mer counting (commonly, 5-mer) within the sequence library can be complemented.

In addition, the percentage of adapters along the sequence reads is usually calculated to determine whether its removal needs to be performed. Usually, adapters in more than 10% of all reads suggest contamination and need to trim (FastQC 2020).

2.2. Read trimming using Trimmomatic

Following the initial RNA-seq quality assessment, trimming steps were implemented through Trimmomatic tool, to remove low score/poor quality reads and adapters inserted for the molecular sequencing.

Trimmomatic is a multithread and pair-aware command line tool optimized to pre-process Illumina NGS data (Bolger, Lohse, and Usadel 2014). It performs various steps for reading filtering and trimming based on their quality classification (Bolger, Lohse, and Usadel 2014). Moreover, the Trimmomatic program identifies and removes adapter sequences previously inserted for the molecular sequencing since they may negatively interfere in downstream analysis. This tool adopts two main strategies to detect and remove technical sequences within the reads: single-end (SE) and paired-end (PE) modes (Bolger, Lohse, and Usadel 2014). The paired-end mode (or *palindrome mode*)

approach is specified for read-through adapter detections, particularly at both reads extremities. Contrarily to the single-end strategy, it does not initiate the alignment at the 5' end (Shrestha and Frith 2013). Instead, it tests for an immediate read-through, with the adapters totally overlapping the start of the opposite reads (Bolger, Lohse, and Usadel 2014). Testing then proceeds in forward and reverse directions, finishing when the overlapping portion no longer reaches into the inserted adapters (Shrestha and Frith 2013). Consequently, the portion of the forward and reverse reads, with no adapters, are reverse-complements. Thus, paired-end sequencing is advantageous to detect not only large contaminant fragments, but also small numbers of adapters bases (Bolger, Lohse, and Usadel 2014). Farther, it facilitates the identification of possible DNA rearrangements, such as insertion-deletion variants or inversions, and splice isoforms in RNA-seq data (Bolger, Lohse, and Usadel 2014).

Figure 10 illustrates the paired-end approach in comparison with the single-end alignment process. Single-end method is represented in steps A and B. Paired-end method is represented in steps C and D.

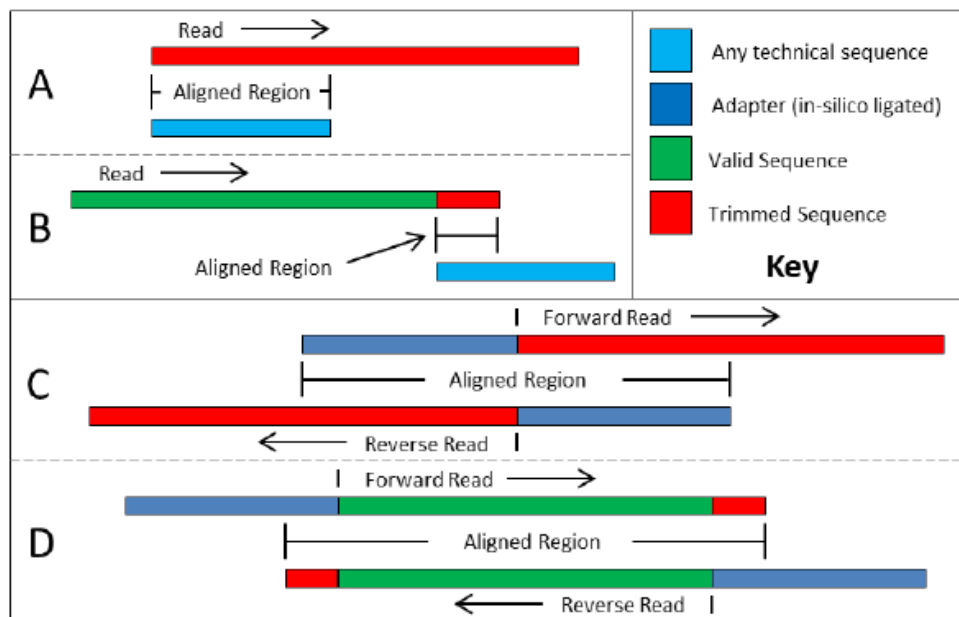


Figure 10 - Representation of both single-end (A and B) and paired-end (C and D) mode strategies. Image retrieved from (Bolger, Lohse, and Usadel 2014).

2.2.1. Implementation of trimming steps

Various trimming parameters are supplied on a command line to run Trimmomatic, usually implemented in a script. Common to both strategies is the initial specification of input and output files, generally in FastQ format, using a classification scheme of phred +33 or phred +64 quality scores (Bolger, Lohse, and Usadel 2014). Single-end mode approach uses only one file as input, generating one file as output. Paired-end sequencing uses two files as input, separating forward and reverse reads. In this approach, four output files are generated:

1. a 'paired' file with forward reads, where both reads survived the processing;
2. a 'paired' file with reverse reads, where both reads survived the processing;
3. a 'unpaired' file with forward reads, where only one read survived the processing;
4. a 'unpaired' file with reverse reads, where only one read survived the processing.

Once specified the input and output files, Trimmomatic tool requires further specification of arguments to perform steps of trimming, filtering, cropping and others. These arguments are individual, added when and/or if needed to the script. As mentioned in the Trimmomatic manual, ILLUMINACLIP is a main argument that allows the location and withdrawal of adapters. To this argument, it is provided a TruSeq file containing the adapters sequences to be assayed. For quality filtering, Trimmomatic offers two separate arguments, namely: SLIDINGWINDOW and MAXINFO. Both methods determine where to cut a read based on the score associated to each base position. The first alternative scans the reads starting at the 5' extremity, cutting and removing the 3' end when it presents groups of bases with quality scores below a specific established threshold. The Maximum Information alternative looks to maintain longer (and moderate) informative reads at a cost of preserve bases with lower quality and errors. Further arguments can be implemented, for instance: removal of reads falling below an established minimal length (MINLEN), sequential low-quality base inquiry and removal (LEADING), set a specific number of bases to remove, regardless of its quality (HEADCROP).

In the present methodological work, the trimming parameters were supplied on a script (*trimmomatic2.sh*), provided and run through MobaXterm terminal. The structure of the Trimmomatic script is illustrated in Figure 11 with the respective selected

parameters and arguments to perform the trimming tasks. Sample C10 was used as an example for demonstration.

```
1 module load Trimmomatic/0.39
2
3 ILLUMINACLIP1="/mnt/scratch/sbism12/scripts/TruSeq3-PE.fa"
4
5 java -jar $TRIMMOMATIC PE -threads 8 -phred33 \
6         #input files
7         "/mnt/scratch/sbism12/data/samples/C10/C10_1.fastq" \
8         "/mnt/scratch/sbism12/data/samples/C10/C10_2.fastq" \
9         #output files
10        "/mnt/scratch/sbism12/analysis/trimmomatic/output/C10_R1_paired" \
11        "/mnt/scratch/sbism12/analysis/trimmomatic/output/C10_R1_unpaired" \
12        "/mnt/scratch/sbism12/analysis/trimmomatic/output/C10_R2_paired" \
13        "/mnt/scratch/sbism12/analysis/trimmomatic/output/C10_R2_unpaired" \
14        ILLUMINACLIP:$ILLUMINACLIP1:2:30:10 LEADING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

Figure 11 – Trimming steps performed using Trimmomatic tool. Overview of the *trimmomatic2.sh* script, and respective parameters specified for the present study. Exemplification using sample C10.

Module load argument allowed the selection of the most recent and available Trimmomatic version (Trimmomatic/0.39). A TruSeq fasta file was supplied to the IlluminaCLIP1 argument, containing the adapters to be trimmed from the reads. Given the selected Trimmomatic approach for adapter sequences removal, one of two distinct TruSeq files can be provided: TruSeq2 and TruSeq3 for single-end and paired-end mode strategies, respectively. In the present work, a TruSeq3 file was specified since a PE sequencing was performed. A classification scheme of phred +33 quality scores was applied.

In this methodology step, two files in FastQ format, separating forward and reverse reads, were used as input for each of the eight samples. Consequently, four output files were generated: 1. two paired files, where both forward and reverse reads survived the processing; 2. two unpaired files, where only one forward and reverse read survived the processing. Both paired files were used in further analyses. Additional arguments, and respective values established as default, were added to the script, namely:

SLIDINGWINDOW for quality filtering, LEADING and MINLEN. Once the trimming analysis was completed, metrics and statistics were retrieved for quality assessment.

2.3. MobaXterm Terminal

MobaXterm (<https://mobaxterm.mobatek.net/>) is a terminal for remote computing in Windows operating system, providing remote network tools (e.g. VNC, X11, FTP) and allowing the connection to diverse embedded servers. Secure shell (SSH) protocol is implemented in this tool to ensure a secure connection and browsing (Ylonen and Lonvick 2006). MobaXterm programs can be extended through plugins (e.g. python, zip, git). This tool was used to connect to the Bioinformatics Hub at Cardiff University, where all the analyses were carried over.

This server has a long-term data storage, and a short-term storage (1 month) with higher speed disks used for active running jobs. Data management, and programming (commonly developed in available scripts) target solely scratch storage (short-term storage). A job to be run is submitted to an existing slurm queue, usually specified in the associated scripts. This available slurm queues are:

1. *defq* (default queue): most commonly used for jobs requiring less running resources, using lower RAM memory and CPU processing;
2. *jumbo*: used for jobs requiring mid-level resources. Suitable for external containers, e.g. 'CharlieCloud' (Priedhorsky and Randles 2017);
3. *mammoth*: suitable for jobs requiring higher memory and processing.

2.4. *de novo* Transcriptome Assembly using Trinity

In this step of Workflow II, a *de novo* transcriptome assembly approach was conducted given that *Carcinus maenas* genome was not yet sequenced. Trinity, a method for the *de novo* reconstruction of full-length transcriptomes, was used. As exhibited in Figure 12, the trimmed samples reads followed two distinct analyses:

1. reconstruction of eight individual *de novo* transcriptome assemblies (Analysis 1);
2. merge of all samples reads, and reconstruction of a single *de novo* transcriptome assembly (Analysis 2).

Analysis 2 was conducted to determine whether the process of concatenation would result in a better *de novo* assembly.

Hereafter, both analyses and approached methodologies will be described in detail. The present approach has also been previously used in the work of (Ferreira et al. 2019).

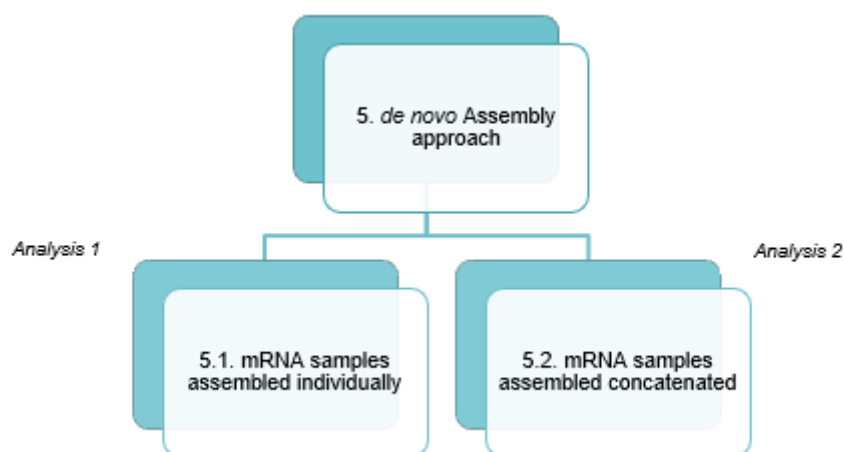


Figure 12 – *de novo* Assembly approach step. Implementation of two distinct analyses: 1. reconstruction of eight individual transcriptomes (*Analysis 1*); 2. concatenation of all samples reads, and reconstruction of a single assembly (*Analysis 2*).

2.4.1 Transcriptome assembly approaches

Genomic and transcriptomic studies have used both model and non-model organisms to reconstruct full-length and high-quality genome and transcriptome sequences, a process designated as an assembly. Two categories of assembly strategies can be distinguished based on the presence or absence of a reference sequence for alignment.

2.4.1.1. Reference-based strategy

A reference-based assembly is conducted for species whose reference genome is known and available (Clarke et al. 2013). Reads relative placement is already established by the template, facilitating its alignment and rearrangement. Once aligned, a layout of reads with overlapping portions is generated, and a final consensus sequence is constructed based on a majority base call process (Wajid and Serpedin

2014). In the case of transcriptomes, RNA-seq short reads are mapped against the template, followed by the detection of overlaps. Overlapping reads are then clustered to build a graph representing all possible isoforms and alternative splicing events (Martin and Wang 2011). This graph is traversed to detect every connecting path and assemble variants. Lastly, path bindings are followed to form isoforms and full and optimized transcript sequences (Martin and Wang 2011). Comparative or population studies often integrate this resequencing approach, assembling small to large genomes and transcriptomes (Lee and Tang 2012). However, transcriptome reference-based assembly is easier to perform for simpler organisms, as these present little alternative splicing and fewer introns, for instance: organisms from the Bacteria and Archaea domains or lower eukaryotic species (Lee and Tang 2012; Martin and Wang 2011). Contrarily, more complex organisms (e.g. plants and mammals) transcriptomes are less accurately assembled from short reads (Martin and Wang 2011).

Though reference sequence-based strategies present multiple advantages, there are also a few corresponding drawbacks. This approach is not possible without a suitable reference genome, with such sequences only being determined and available for a very small portion of organisms (Haas et al. 2013). Reference genomes from closely related species are commonly used to outline this problem. However, not all reads map to the template without mismatches or gaps, with consequent miss of transcripts from divergent genomic regions (Voshall and Moriyama 2018; Martin and Wang 2011). Additionally, this alternative is hampered by rearrangements and differential order of some of the contigs among the reference and the target genomes (Lee and Tang 2012).

It is also perceptible how the quality of the reference genome impacts the read mapping quality, and the entire assembly process (Voshall and Moriyama 2018). For instance, when genes of interest possess gaps in the genome assembly, all or part of their fragments will not be assembled into the final transcript sequence (Voshall and Moriyama 2018). Another example is the presence of misassemblies and large genomic deletions in many organisms' genome assemblies, leading to partially or poorly reconstructed transcriptomes (Martin and Wang 2011). Ambiguity may also be an issue if reads map to multiple locations within a genome (Voshall and Moriyama 2018). In the antecedent sequencing process and during splicing events, errors introduced by short-read aligners are likewise the source for final gapped and highly fragmented assemblies (Grabherr et al. 2011).

Nevertheless, these strategies are still very used when high-quality reference genomes are available, given their accuracy and sensitivity in assembling full-length transcripts. Compared to other approaches, their lower complexity in transcriptome

assembly and less computational requirements facilitate its application in multiple studies (Voshall and Moriyama 2018; Huang, Chen, and Armbruster 2016). Further, if combined with gene predictions, reference-based assembly strategies are effective for comprehensive transcriptome annotation (Martin and Wang 2011). Various algorithms and software use and integrate this approach, for instance: Cufflinks, Scripture, and others (Trapnell et al. 2010; Guttman et al. 2010).

2.4.1.2. *de novo* assembly strategy

A *de novo* assembly is particularly suitable for non-model organisms with no reference genome, generating full-length transcripts solely through short RNA-Seq reads (Voshall and Moriyama 2018; Huang, Chen, and Armbruster 2016). Withal, it is also recurrently conducted for species with poor-quality or incomplete genome sequences (Huang, Chen, and Armbruster 2016; Grabherr et al. 2011). Its process seeks to construct sets of contig sequences, each generated from overlapping oriented reads. Afterwards these contigs are assembled into scaffolds (Lischer and Shimizu 2017). A scaffold may contain more than two contigs by binding more to its extremities (He et al. 2013). Lastly, groups of scaffold sequences are computed into final transcripts. When compared to the reference-based approach, this strategy presents several advantages. It does not depend on a reference template. Even when one is available, the *de novo* assembly allows for the recovery of novel transcripts transcribed from segments missing in the reference genome (Martin and Wang 2011). Another beneficial aspect is the no need for correct alignment of reads to discover splice variants per gene not annotated in the reference sequence (Martin and Wang 2011; Lu, Zeng, and Shi 2013). Moreover, it easily detects transcripts from unknown sources. This and more aspects of the *de novo* approach have a positive impact in many biological analyses, for instance: evolutionary and gene discovery studies, facilitating quantitative assessments of differential gene expression (Smith-Unna et al. 2016). On the other hand, a few disadvantages are also associated with this method. *de novo* approaches require broad computing resources to assemble transcriptomes from larger organisms and high sequencing depth to reconstruct full-length transcripts. They are also susceptible to errors introduced during sequencing, easily affecting the assembly process (Cocquet et al. 2006). Additionally, this approach hardly differentiates highly similar transcript sequences, assembling them into a single transcript (Martin and Wang 2011). Compared with resequencing strategies, the *de novo* strategy tends to produce highly fragmented assemblies. In an attempt to outline this problem, genomic resources from closely related species are commonly used to guide the transcriptome

reconstruction for non-model organisms (Huang, Chen, and Armbruster 2016). However, overcoming assembly fragmentation has shown to be difficult. Therefore, existing assemblers have been encouraged to evolve their base assembly process by adopting new computational solutions and schemes. Currently, multiple approaches are optimized to perform an assembly with no reference sequence, for instance: the greedy approach, the exhaustive approach or the overlap-layout-consensus strategy (Lee and Tang 2012). The latter has been the most used, integrating concepts of graph theory, such as overlap graphs, bidirected graphs and de Bruijn graphs. Assemblers using this approach are categorized based on the graph constructed in the overlap phase (Bowe et al. 2012). These diverse computational associations look to solve *de novo* assembly problems, as well as improve and facilitate its process.

2.4.1.2.1. Overlap-layout-consensus approach

Overlap-layout-consensus is a three-step approach complemented with graph-based schemes to optimize the assembly process (Halim 2020). As a primary step, an overlap graph is constructed through the combination of all pairs of sequence reads with respective best overlapping reads (Lee and Tang 2012). In this graph structure, reads are represented by nodes, and bindings among every pair of overlapping reads are represented by edges. Secondly, all reads are orderly traversed, forming multiple connecting paths (He et al. 2013). This structure is designated as a graph layout (Wajid and Serpedin 2014). A smaller number of identified paths suggests a less fragmented assembly (Lee and Tang 2012). Ideally, it is intended to find a single path that sequentially visits all nodes exactly one time, finishing at the node where it started. This process is called the Hamilton cycle or circuit (Compeau, Pevzner, and Tesler 2011). Lastly, a consensus step determines which bases from the overlapping groups of reads should be present at a particular location in the reconstructed genome or transcriptome (Wajid and Serpedin 2012). This approach faces the obstacle of finding only one path, binding all nodes from root to child node (Wajid and Serpedin 2014). Most commonly, several disjointed graphs are extracted, each depicting a contig (Wajid and Serpedin 2012). The branches of each smaller graph compete to define the final contig, with the shortest being discarded. Posteriorly, these graphs contigs are sequentially connected, a process designated as scaffolding (Wajid and Serpedin 2012). Lastly, a final genome or transcriptome is generated. Assembler software integrating the overlap-layout-consensus approach, complemented with overlap graphs, were developed in the late 1990s and early 2000s, for instance: Newbler, Mira, and others (Clarke et al. 2013; Chevreux et al. 2004; Margulies et al. 2005).

Most recently, assemblers implementing overlap graphs have been facing problems with the vast amount of data produced by HTS technologies. Since billions of short sequencing reads are generated daily, it becomes computationally hard to compare all pairs of reads and determine which should be joined into contigs (Grabherr et al. 2011; Bowe et al. 2012). Further, differences among read lengths affect overlaps' detection using these graphs concepts (Bowe et al. 2012) solely. This way, improved graph-based schemes (e.g. de Bruijn graphs) have been replacing the overlap graphs, in the assemblers overlap phase (Bowe et al. 2012; Li et al. 2010).

2.4.1.2.2. de Bruijn graph concept

A de Bruijn graph is a computational technique that computes a k -mer to each of its nodes, usually with a length of 5 bp (Chikhi et al. 2014). Connections between two nodes are established when a $k-1$ overlap occurs among the corresponding short reads (Bowe et al. 2012; Chikhi et al. 2014). As previously mentioned, these bindings, represented by arrows, are designated as edges (Compeau, Pevzner, and Tesler 2011). When computing an assembly, each sequential node's k -mer is dislocated by one position, aligning a final transcriptome (Compeau, Pevzner, and Tesler 2011). The edges course and associated paths provide an insight of possible full-reconstructed transcripts. However, as the overlap graphs, the de Bruijn approach faces the computational obstacle of visit each graph node only once and extracting a single connecting path (Compeau, Pevzner, and Tesler 2011). As an alternative, the focus has been on reconstructing de Bruijn graphs structure, by integrating a cycle that goes through all graph edges, one time. This method is known as the 'Eulerian cycle' or circuit (Idury and Waterman 1995; Compeau, Pevzner, and Tesler 2011). Inversely to the Hamiltonian strategy, this cycle computes a k -mer to each edge and $(k-1)$ -mers to each graph node. When reconstructing an assembly, this cycle dislocates each sequential k -mer by one position, in this case, from successive edges (Compeau, Pevzner, and Tesler 2011). With these adjustments, it was intended for de Bruijn graphs to present more memory efficiency and large-scale transcriptome assembly improvement. Figure 13 illustrates a simplified representation of de the Bruijn graph strategy, considering both Hamilton and Eulerian cycles.

As with any other method, there are still a few challenges to surpass when using the de Bruijn graph. The main problem when assembly transcriptomes is the loss of information when short-reads are split into k -mers (Clarke et al. 2013). For example, reads from two genes sharing a sequence with a length higher than k would be

incorrectly joined by this approach (Clarke et al. 2013). Nonetheless, modern algorithms currently apply the de Bruijn strategy since its graphical computational method, implementing Eulerian cycles, allows for optimising the assembly process by converting it into a tractable computational problem (Compeau, Pevzner, and Tesler 2011). Velvet-Oases and Trinity are assembler software successfully implementing de Bruijn approach, directly traversing its graph to assemble each possible isoform (Martin and Wang 2011; Grabherr et al. 2011; Schulz et al. 2012). ABySS and SOAPdenovo were also reported to properly compute this strategy in the transcriptome assembly of multiple organisms (Li et al. 2009; Birol et al. 2009; Luo et al. 2012). The de Bruijn graph application is also favoured since the number of nodes in its graph is proportional to the genome or transcriptome size (Lee and Tang 2012). Moreover, in cases of high coverage and low error rate, the de Bruijn approach allows finding contigs by traversing highly covered edges in its graph, unlikely to contain sequencing errors (Lee and Tang 2012).

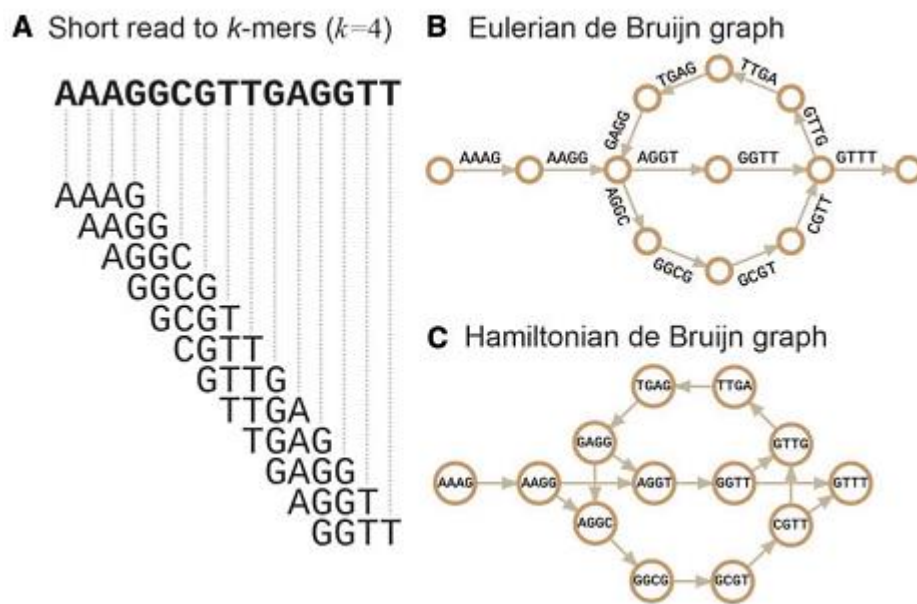


Figure 13 – de Bruijn graph approach. In stage A, short-reads are partitioned into k -mers with a length of 4 bp. Stage B illustrates the Eulerian cycle, where a k -mer is computed to each edge. Stage C illustrates de Hamilton cycle, where a k -mer is computed to each node. Image retrieved from (Sohn and Nam 2018).

2.4.2. Trinity: a method for *de novo* transcriptome assembly

Trinity is a short-read RNA-seq assembler released by (Grabherr et al. 2011), developed to recover full-length transcripts and isoforms when no reference transcriptome is available. Compared to alternative methods, this *de novo* assembler is distinguished, given its effectiveness and sensitivity during transcriptome reconstruction (Zhao et al. 2011). Trinity's assembly workflow consists of three software modules: Inchworm, Chrysalis and Butterfly. These are schematized in Figure 14.

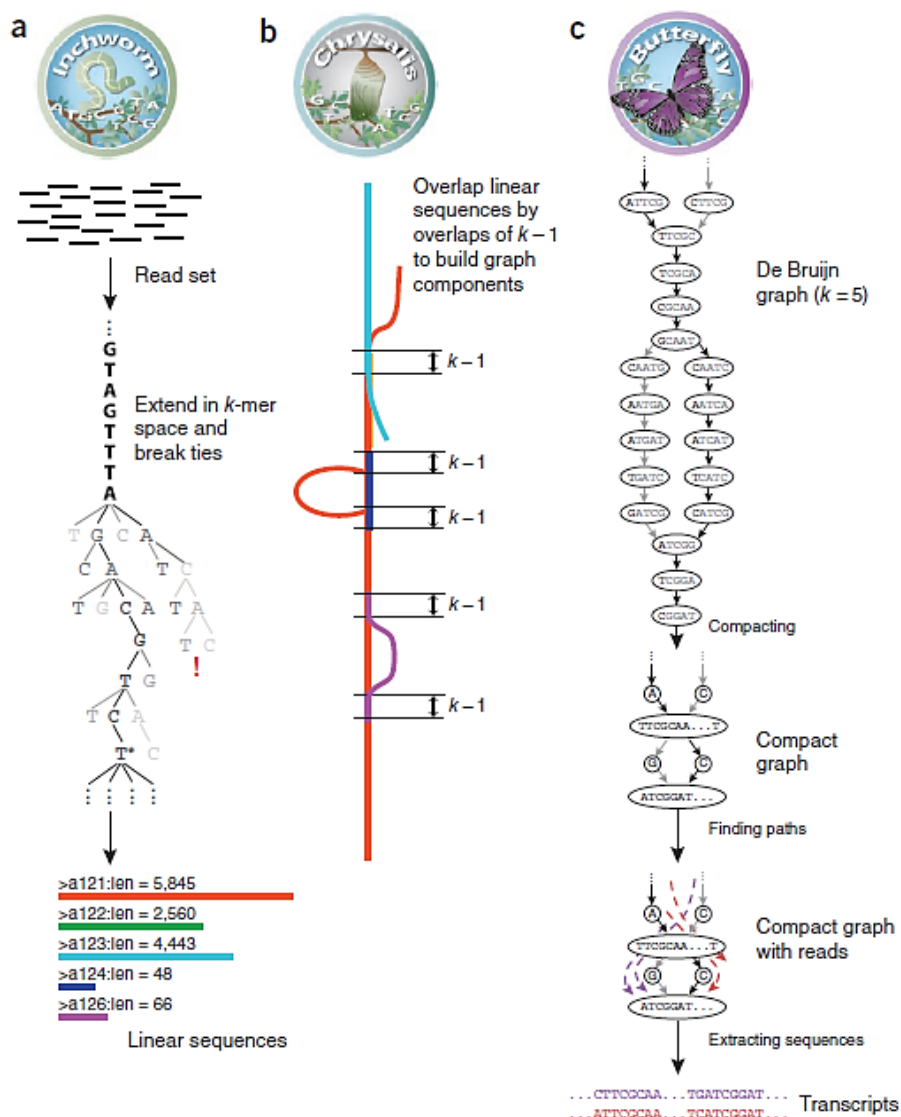


Figure 14 - Overview of Trinity modules: Inchworm (a), Chrysalis (b) and Butterfly (c). Image retrieved from (Grabherr et al. 2011).

Inchworm is the first step in Trinity's pipeline, consisting in defining k -mers (usually, $k=25$), from all available sequence reads, based on its frequency of appearance: it excludes low-quality k -mers with single occurrences, keeping the most common and recurrent (Grabherr et al. 2011). It finds the most frequent k -mers through a $(k-1)$ -mer overlapping system, followed by a dual-direction and greedy extension process to generate linear transcript contigs (Haas et al. 2013). This step easily operates in transcriptomes, given its lack of repetitive sequences. However, it limits the account of alternatively spliced transcripts, reporting only unique portions (Haas et al. 2013). Figure 14 represents the Inchworm workflow (left side) implemented by the Trinity assembler.

After contigs are generated, Chrysalis step clusters related contigs into connected components (Haas et al. 2013). This aggregation occurs when contigs share at least one overlapping $(k-1)$ -mer, and reads cover the union between contigs (Grabherr et al. 2011). Then, it builds a de Bruijn graph for each formed component, with word sizes of $k-1$ and k computed to its nodes and edges, respectively. Lastly, each read is assigned to a particular component based on the highest number of shared k -mers (Grabherr et al. 2011). Figure 14 represents Chrysalis workflow (middle), implemented by Trinity assembler.

Once de Bruijn graphs are built for each group of contigs, the Butterfly step starts processing each individual graph in parallel (Haas et al. 2013). Then, it compacts all constructed graphs, tracing all its supporting RNA-seq reads and established paths. If bindings are not verified among reads in a particular graph, this step divides it into disconnected sub-graphs, analysing each separately (Haas et al. 2013). Contrarily, when potential paths are identified, Butterfly uses a traverse method to cross them while keeping its supporting reads (Grabherr et al. 2011). Finally, it outputs and reports all plausible linear transcripts. This last computing stage is advantageous for resolving alternatively splice problems in isoforms and paralogous gene transcripts (Grabherr et al. 2011). Figure 14 represents Butterfly workflow (right side), implemented by Trinity assembler.

2.4.2.1. Trinity implementation and requirements for assembly

Trinity software is executed from a command line, usually via a single Perl script, referred to as `trinity.pl` (Henschel et al. 2012). As input, it requires files with paired-end or single-end short-read RNA data in FASTA or FASTQ formats. In the case of paired-end sequencing, when multiple runs are being conducted, these reads may be

concatenated into two files, separating forward and reverse reads. As the final output, Trinity generates a single FASTA file per run containing the reconstructed transcript sequences. Moreover, this tool allows the incorporation of external software as plug-ins, for instance: Samtools (Danecek et al. 2021), Salmon (Patro et al. 2017) and Bowtie2 (Langmead and Salzberg 2012). It also contains third-party software components, such as Trimmomatic (Bolger, Lohse, and Usadel 2014), Jellyfish (Marçais and Kingsford 2011) and Collectl (Oberman and Seger 2007).

2.4.3. Reconstruction of eight *de novo* transcriptome assemblies

As previously mentioned, Trimming performance generated two paired files for each sample, where both forward and reverse reads survived the processing. During Analysis 1 of the assembly approach step, these files were used as input to reconstruct eight full-length transcriptomes. The *de novo* assemblies were performed through a script (*Trinityrnaseq_266_ysgo.sh*), provided and run through MobaXterm. Figure 15 illustrates the structure of Trinity script, and respective selected arguments and parameters. Sample C10 was used as an example for demonstration.

```
1 module load python/3.9.7
2 module load bowtie2/v2.4.1
3 module load samtools/1.13
4 module load trinityrnaseq/Trinity-v2.6.6
5 module load salmon/v0.9.1
6
7
8 time Trinity --seqType fq \
9             --CPU $SLURM_NTASKS \
10             --max_memory 240G \
11             --normalize_reads \
12             #input files
13             --left "/mnt/scratch/sbism12/analysis/trimmomatic/output/C10_R1_paired" \
14             --right "/mnt/scratch/sbism12/analysis/trimmomatic/output/C10_R2_paired" \
15             #output file
16             --output "/mnt/scratch/sbism12/analysis/trinity_analysis1/C10" \
17
```

Figure 15 –Overview of the *Trinityrnaseq_266_ysgo.sh* script, and respective parameters specified for the present study. Exemplification using sample C10.

Module load argument allowed the selection of multiple modules (python, bowtie2, samtools, trinityrnaseq/Trinity and salmon) in its most recent and available versions, needed for Trinity assembler execution (Langmead and Salzberg 2012; Danecek et al. 2021; Patro et al. 2017; Van Rossum and Drake Jr 1995; Grabherr et al. 2011). Since millions of paired-end sequences were to be assembled, Trinity's in silico normalization was applied upon the reads (*normalize_reads*) to improve assembly efficiency (Brown et al. 2012). FastQ (*fq*) format was specified. In Figure 15, it is demonstrated the selection of both input paired files for sample C10. Designations R1_paired (left) and R2_paired (right) remit to forward and reverse trimmed reads, respectively. An output folder containing a single FASTA file was generated for each sample.

2.4.4. Reconstruction of a single *de novo* transcriptome assembly

During analysis of the assembly approach step 2, all sample reads were concatenated into two fasta files, separating both trimmed forward and reverse reads. The concatenation was achieved through Linux/Unix-like operating system *cat* command, jointly with the redirection operator '>'. *cat* command allows, besides concatenating files, various other processes, for instance: creating single or multiple files, view content of a file, display tab separate lines in files, sorting contents of multiple files in a single file, and others. This applied operators conjunction is represented below:

1. mRNA forward reads concatenation: `cat 'C10_R1_paired' 'C13_R1_paired' 'C3_R1_paired' 'G1R1_R1_paired' 'R11_R1_paired' 'R12_R1_paired' 'R15_R1_paired' 'R3_R1_paired' > 'R1_paired_concatenated'` (output file);
2. mRNA reverse reads concatenation: `cat 'C10_R2_paired' 'C13_R2_paired' 'C3_R2_paired' 'G1R1_R2_paired' 'R11_R2_paired' 'R12_R2_paired' 'R15_R2_paired' 'R3_R2_paired' > 'R2_paired_concatenated'` (output file).

The single *de novo* assembly was also performed through the *Trinityrnaseq_266_ysgo.sh* script. As illustrated in Figure 16, most parameters and arguments remained unaltered from the first Trinity assembler run. The 'R1_paired_concatenated' and the 'R2_paired_concatenated' files were specified as input. An output folder containing a single FASTA file was obtained.

```

1 module load python/3.9.7
2 module load bowtie2/v2.4.1
3 module load samtools/1.13
4 module load trinityrnaseq/Trinity-v2.6.6
5 module load salmon/v0.9.1
6
7
8 time Trinity --seqType fq \
9             --CPU $SLURM_NTASKS \
10            --max_memory 240G \
11            --normalize_reads \
12            #input files
13            --left "/mnt/scratch/sbism12/analysis/R1_paired_concatenated" \
14            --right "/mnt/scratch/sbism12/analysis/R2_paired_concatenated" \
15            #output file
16            --output "/mnt/scratch/sbism12/analysis/trinity_analysis2/reads_concatenated" \
17

```

Figure 16 –Overview of the *Trinityrnaseq_266_ysgo.sh* script, and respective parameters specified for the present study.

2.4.5. Assess the overall quality of the assemblies

For both conducted analyses, a supplementary script - *TrinityStats.sh* - was provided and run through MobaXterm, to quantitatively and qualitatively assess the overall quality of the reconstructed assemblies. Figures 17 and 18 illustrate the structure of *TrinityStats.sh* script for Analyses 1 and 2, respectively. The FASTA files generated as output once concluded the assemblies reconstruction, were used as input. In Figure 17, sample C10 was considered for exemplification.

```

1 module load bowtie2/v2.4.1
2 module load samtools/1.10
3 module load trinityrnaseq/Trinity-v2.6.6
4
5 TrinityStats.pl "/mnt/scratch/sbism12/analysis/trinity_analysis1/C10/Trinity.fasta" > / #input
6 file           "/mnt/scratch/sbism12/statistics/trinity_analysis1/C10" #output file

```

Figure 17 – Overview of the *TrinityStats.sh* script for Analysis 1.

```

1 module load bowtie2/v2.4.1
2 module load samtools/1.10
3 module load trinityrnaseq/Trinity-v2.6.6
4
5 TrinityStats.pl "/mnt/scratch/sbism12/analysis/trinity_analysis2/concatenated_reads/Trinity.fasta" > / #input
6 file           "/mnt/scratch/sbism12/statistics/trinity_analysis2/concatenated_reads" #output file

```

Figure 18 – Overview of the *TrinityStats.sh* script for Analysis 2.

A few straightforward and basic statistics were retrieved for evaluation, namely:

1. number of total assembled transcripts;
2. number of total read count;
3. the contig average and median length;
4. percentage of GC;
5. metrics N10 to N50.

Assessing genome and transcriptome assemblies, compared to known and established standards, is required to prevent negative impact in downstream analysis (Alkan, Sajjadian, and Eichler 2011). Though it is unlikely to obtain a perfect assembly with no errors, using any assembly approach, these can be minimized or even rectified (Alhakami, Mirebrahim, and Lonardi 2017). In the case of *de novo* transcriptome reconstructions, with no known reference available, there is an urgent need to assess assemblies quality and identify the set of well-assembled contigs to prevent negative influences in biological interpretations (Smith-Unna et al. 2016). Several factors affect the accuracy and efficiency of a transcriptome assembly process, for instance: prior errors occurring during sequencing, variability and mutations, or defective coverage of transcripts (Smith-Unna et al. 2016). Algorithmic simplifications are also the source of assembly artefacts (Smith-Unna et al. 2016). As previously exemplified, partitioning reads into shorter words can lead to substantial information loss. To optimize and improve an assembly, this and other computational adjustments may do the opposite by raising more errors and problems to the process. Deliberately allowing faults to mitigate the complexity of the assembly problem is also a major reason behind high transcriptome fragmentation, local misassembly of contigs, and blending of transcripts (Smith-Unna et al. 2016). An example of allowance is to decide whether to maximize

contigs and scaffolds length, at the expense of possibly induce misassemblies, or vice versa (Alhakami, Mirebrahim, and Lonardi 2017).

A wide range of statistics is commonly used to evaluate the quality and performance of transcriptome assemblies. The most straightforward and basic metrics are based on the number and length of the generated sequences (Voshall and Moriyama 2018). For instance:

1. minimum and maximum sequences length (bp): the length of the shortest and longest sequence, respectively;
2. contigs average length (bp): the mean length of the produced contigs;
3. contigs median length (bp): the length where half the sequences are longer, and half are shorter;
4. number of total assembled transcripts.

Once concluded an assembly, it is considered better to obtain longer and fewer contigs (Wajid and Serpedin 2014). Contrarily, a high number of incompletely assembled transcripts (or with low coverage) is not expected, since they are more likely to present contigs with low scores due to fragmentation and gapped regions (Smith-Unna et al. 2016). The N50 is another frequently retrieved statistic to assess the contiguity of an assembly. It is defined by the length of the shortest contig for which equal and longer length contigs cover 50% or more of the assembly (Alhakami, Mirebrahim, and Lonardi 2017). This metric can be extended to provide other thresholds (e.g. N40 to N10) with similar meanings. As the N-value decreases, the corresponding length generally increases (Alhakami, Mirebrahim, and Lonardi 2017). Particularly for genome assemblies, where the target number of sequences is known, higher N50 values indicate a better assembly quality (Wajid and Serpedin 2014; Voshall and Moriyama 2018). This metric may be artificially increased by merging contigs but at the expense of raising the number of misassemblies (Gurevich et al. 2013). NG50 resembles N50, except it covers the genome size instead of the assembly size (Alhakami, Mirebrahim, and Lonardi 2017). More specifically, it retrieves the length of the scaffold at which 50% of the genome length is covered (Wajid and Serpedin 2014). These initial metrics are commonly complemented with an evaluation of the percentage of GC in the assembly. This statistic (GC%) is indicative of the total number of G and C nucleotides divided by the total length of the assembly (Gurevich et al. 2013).

Though contig length estimates are the primary to be analysed, they do not necessarily reflect the quality of transcriptome assemblies. A reliable quality measurement cannot be retrieved, with assembled contigs length fluctuating according to the lengths of the transcripts being assembled, or longer contigs being underlining, leading to over-assembled sequences by numerous assemblers (Voshall and Moriyama 2018). Metrics based on coded protein similarity have been proposed in (Voshall and Moriyama 2018) to improve the assessment of transcriptome assemblies. Usually, this is performed through similarity searches against annotated databases or lineage specific conserved single-copy protein databases, e.g. BUSCO (Simão et al. 2015). When comparing an organism to a closely related species with a well-defined and available transcriptome, most contigs correctly assembled presumably have homologs in the database. If a large percentage of contigs matches a high number of similar proteins, it is presumable that the sequences are accurately assembled. Statistics based on benchmark transcriptomes are also suggested by (Voshall and Moriyama 2018) to evaluate transcriptome assemblies. This strategy compares the assembly output against correct sequences of all known transcripts and expression patterns. Various other metrics have been implemented in multiple quality assessment tools, for instance: percentage of misassemblies and structural variations, number of genes in an assembly, number of present gaps, scaffold statistics, and others (Gurevich et al. 2013; Wajid and Serpedin 2014).

2.4.5.1. Computational tools for data analysis and visualization

Additionally, computational analyses (PCA and MA) with correspondent plots, and a Heatmap, were conducted and obtained through OmicsBox to complement the previously retrieved statistics (BioBam 2019). Once concluded the analysis, OmicsBox generated the following reference: “OmicsBox – Bioinformatics Made Easy, BioBam Bioinformatics, March 3, 2019, <https://www.biobam.com/omicsbox>”.

Principal Component Analysis (PCA) is a multivariate mathematical algorithm that reduces the dimensionality of a matrix table (or a data set) while retaining most of its variation, and computes the majority of its significant information into a set of new orthogonal variables, known as principal components (Abdi and Williams 2010; Jolliffe and Cadima 2016). This data table represents its features through dependent variables, commonly inter-correlated (Abdi and Williams 2010). Principal components (PCs) are uncorrelated linear functions (or combinations) of the initial observed variables, generally obtained through the appliance of a matrix factorization method

denominated singular value decomposition, or SVD (Lall et al. 2018; Vezzi, Narzisi, and Mishra 2012). In other words, determined by solving an eigenvalue/eigenvector problem, as mentioned in Jolliffe et al. 2016 (Jolliffe and Cadima 2016). There are two main PCs describing the largest portion of the variance of the data set, with the second component being orthogonal to the first one (Abdi and Williams 2010). In other words, the direction uncorrelated to the first component, along which the samples demonstrate a maximal variation (Ringnér 2008). All remaining PCs are computed similarly. PCA technique allows samples to be plotted and their distribution visualized. This way, it is possible to observe similarities and differences among the data, identify if samples may be grouped and discover potential clusters already formed (Ringnér 2008).

Regarding RNA-seq data sets, PCA is widely used to visualize genes and detect differential gene expression. Its capability in dimensionality reduction and representation with smaller groups of variables is advantageous when considering transcriptomes since they produce large data sets and endue a significant number of genes (Lall et al. 2018). One example of the applicability of PCA in RNA transcripts sets is the work developed in (Ringnér 2008). This research underlines the ability of PCA to reduce a set of genes down to two dimensions. It also suggests this technique as a useful first step before sample clustering and classification, given the uncorrelation among the principal components. Additionally, the implementation of SVD during PCA allows identifying and filtering artefacts patterns (interfering in data variation), estimation of missing data and association of genes with expression patterns (Ringnér 2008; Nielsen et al. 2002). Lastly, it mentions the disadvantage in discovering unknown groups using PCA. Another example is the study conducted in (Vezzi, Narzisi, and Mishra 2012). It demonstrates how this technique selects key attributes of the feature space without loss of information and determines the accuracy of evaluator metrics to describe the quality of transcriptome assemblies.

In computational biology, one common comparison among two treatment conditions is the log fold-change versus normalized mean expression (or counts). MA plots, also known as static scatter plots, are commonly used to display this type of comparison (Sheharyar et al. 2020). Base-2 log fold-change is generally represented in its y-axis, while the normalized mean expression is along its x-axis (McDermaid et al. 2019). Data is visualized as points in these plots. Since log fold-changes imply standard thresholds, MA plots address data exhibiting below-threshold adjusted p-values by highlighting them as coloured points (McDermaid et al. 2019).

MA plots are frequently applied in computational analyses of raw RNA-seq reads (e.g. transcriptomic assembly approaches), particularly when considering genes with distinct expression levels across two or more conditions (Oshlack, Robinson, and Young 2010; Pertea et al. 2016). These genes are often referred to as differentially expressed genes (or DEGs). With an established threshold (generally set as value 1) and considering a data point as a single gene, MA plots infer genes being upregulated when most points fall above the defined threshold on the y-axis. The opposite occurs for downregulated genes. Since most data is usually distributed close to the log fold-change axis, with a tendency to thicken along the normalized mean counts axis (in other words, with similar expression pattern), extreme values along the y-axis are interpreted as genes with high levels of differential expression (McDermaid et al. 2019). Thus, when one or more samples are evidenced with a drastically different distribution from the others, bias identification and quality assessments should be performed on the studied data.

Venn diagrams, also known as set or logic diagrams, are graphical representations of relations among two or more groups of elements. Its illustrations occur in the form of circles or ellipses, overlapping when elements are common to more than one set (Bardou et al. 2014). In bioinformatics, these diagrams are widely used to compare subsets, for example, when considering pairwise differential expression analysis. In other words, to visualize genes with distinct expression patterns and how they connect (Sun et al. 2019). When considering RNA-seq data and differentially expressed genes, Venn diagrams are usually complemented with heatmaps. These interactive visualizing methods are displayed in a grid format, with each row representing a gene and each column representing a sample. Each grid box is coloured with alternating intensities, indicating differences in gene expression. In biological research, the two colours most commonly used in this type of representation are red and green, associated with upregulated and downregulated genes, respectively. Heatmaps may also be combined to clustering methods, in order to group samples and/or genes, and identify possible associations or similarities of expression (Grant, Manduchi, and Stoeckert Jr 2007).

2.5. Assessment of Redundancy and Completeness

A redundancy and completeness evaluation of the reconstructed transcriptome assemblies was conducted implementing EvidentialGene (or Evigene) software and BUSCO tool, respectively. As exhibited in Figure 19, Analysis 1 was subdivided into:

- a) Analysis 1.1. All assemblies were assessed individually;
- b) Analysis 1.2. The assemblies were concatenated into one fasta file and posteriorly assessed.

Analysis 1.2 was conducted to determine whether the process of concatenation would result in a better *de novo* assembly. Analysis 2 proceeded in the single reconstructed *de novo* assembly.

Hereafter, the approached methodologies will be described in detail.

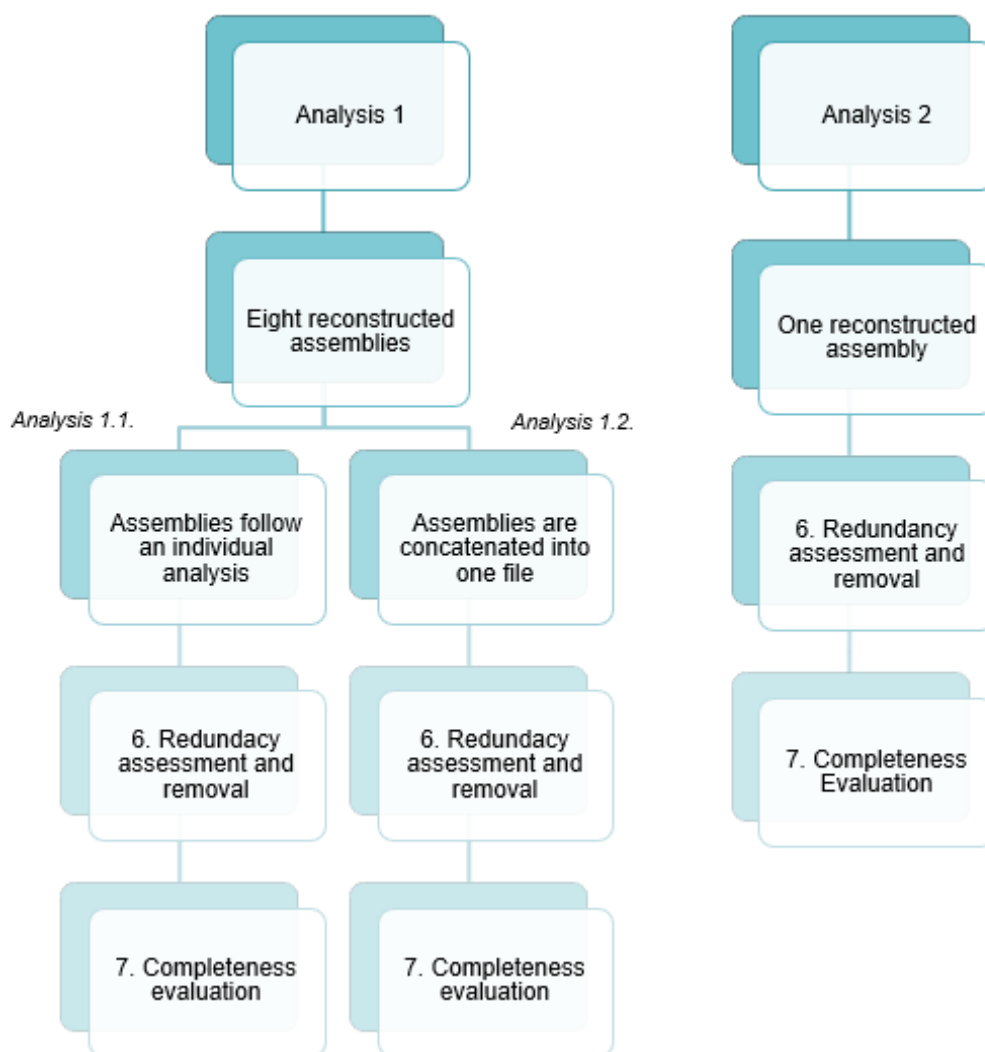


Figure 19 - Statistical computation and assessment steps for both Analyses 1 and 2. Evaluation of redundancy (EvidentialGene) and completeness (BUSCO tool). Representation of Analysis 1 partition into Analysis 1.1. and Analysis 1.2.

2.5.1. Evaluation and removal of redundancy

When assembling a transcriptome, it is challenging to accurately reconstruct and distinguish highly similar transcripts expressed from orthologs and paralog genes (Nakasugi et al. 2014). This becomes hardened with transcript isoforms (Nakasugi et al. 2014). For a brief and simplified contextualization, transcript isoforms arise from alternative splicing (Haas and Zody 2010). Messenger RNA molecules with distinct sequential exons are produced from a single gene, expanding the number of protein products. Orthologs and paralog genes are two types of homologs genes, that is, genes arising from a common DNA ancestral sequence (Koonin 2005). Orthologs genes are derived from a speciation event and are found in different organisms with similar functions (Dufayard et al. 2005; Koonin 2005). Contrarily, paralog genes diverged from one another within a species, commonly developing new functions through mutational events (Dufayard et al. 2005; Koonin 2005). Genes ending one or more homologs in a genome with a resemble biochemical function are referred to as redundant genes (Peng 2019). From an evolutionary perspective, gene redundancy benefits the survival of an organism when one copy of the homologs becomes non-functional or defective (Peng 2019). However, its presence may significantly interfere in downstream analysis, for instance, assembly processes. In *de novo* assembly, redundancy commonly occurs when sets of contigs exhibit homology to one gene, leading to a number of estimated genes significantly lower than the number of contig sequences (Ono et al. 2015). These over-assemblies are usually induced when reconstructions are based on insufficient and incomplete reads (Ono et al. 2015). In the case of transcriptome assembly, redundant transcripts pose various difficulties, for instance, correctly identify and detect differentially expressed genes (Duan et al. 2012). This way, a considerable number of programs and softwares have been developed to minimize redundancy and non-redundant sequences, for instance: CD-HIT (Li and Godzik 2006), TransRate (Smith-Unna et al. 2016) and EvidentialGene (Gilbert 2016b). CD-HIT, program introduced by (Li and Godzik 2006) in the early 2000's, is widely used for clustering biological sequences and determining its similarity without performing a sequence alignment. Its algorithm is based on short-word filtering, and was firstly limited to remove redundancy from protein databases (Li, Jaroszewski, and Godzik 2001). Currently, DNA and RNA sequences can also be analysed and compared through an extension of this software known as CD-HIT-EST (Li and Godzik 2006). In summary and as described in (Li and Godzik 2006), this greedy-incremental clustering programs sort sequences in order of decreasing length. Each sequence is

analysed individually, with the longest sequences becoming cluster representatives. All remaining and smaller sequences are then compared to any first cluster sequences. During the comparison, if the similarity between a sequence and a representative is above an established threshold, that sequence is assigned to the cluster of that representative. When the opposite happens, a new cluster is defined with that sequence as the representative. Short-word filtering is the process confirming whether similarity is or not below the clustering threshold. This prevents performing a sequence alignment. However, one must be executed if similarity cannot be verified and confirmed. The introduction of these software allowed an effective removal of redundancy and an improvement in homology detection, for instance, among *de novo* contig assemblies (Ono et al. 2015). Another method proposed for assembly filtering and optimization is TransRate, introduced by (Smith-Unna et al. 2016). To assess *de novo* transcriptome assemblies accuracy and completeness, this reference-free method provides measures at a contig (individually) and an assembly level. In other words, it only uses input reads for assessment. Briefly and as described in (Smith-Unna et al. 2016), reads are mapped to the assembled contigs. Following, multimapping reads are assigned to their contig of origin, and contig score components are collected. This allows calculating and determining contig scores, and consequently, the assembly score. The strategy of read mapping data has been suitable to detect common errors (e.g. fragmentation, misassemblies, redundancy) in assemblies of all types of RNA. Evidently, while filtering redundant sequences, clustering and read mapping strategies retrieve best groups of longest and better expressed transcripts, respectively. Though these demonstrate capability to recover improved transcript sets, its accuracy is feeble when compared to methods based on coding sequences (CDS) measurements.

2.5.1.1. EvidentialGene: a tool for transcript and gene sets reconstruction and redundancy removal

EvidentialGene (<http://arthropods.eugen.es.org/EvidentialGene/>) is a software that looks to accurately construct complete gene sets or transcriptomes for model and well-known eukaryotic organisms (e.g. animals and plants), based on CDS and protein length (Nakasugi et al. 2014; Gilbert 2016b). This genome project, developed by Don Gilbert at Indiana University, focuses on the coding potential to select the 'best' set of putative transcripts and, consequently, built genes with more precision and efficiency (Nakasugi et al. 2014). This approach surpasses the single use of gene prediction (a

process of locating gene coding regions along a genome) and gene assembly from RNA for biological validity, improving gene annotation and identification of gene function (Gilbert 2010; Gilbert 2016a; Gilbert 2016b). Instead, EvidentialGene combines both methods and applies gene modelling to extract the best of their results (Gilbert 2016a). Briefly and as described in (Gilbert 2016a), its main workflow steps involve: 1. producing various predictions and transcript assembly sets with associated quality models; 2. annotating models with evidence (e.g. CDS, exons and introns, protein homology, and others); 3. attribute scores to these models, and remove those below a minimum established evidence score; 4. select models with highest punctuations; 5. retrieve a final best (and close to perfect) set. Multiple studies have been demonstrating that emphasising this CDS approach and associated methodology provides consistency with gene biology, as well as classifications beyond expression level and length, not discarding alternates and paralogs, as it may happen with other strategies (Gilbert 2019b). Moreover, it prevents unreliability in gene sets, commonly caused by sequencing and assembly errors, incorrect modelling of complex genes, and errors with origin in public databases (Gilbert 2019a). When recovering transcripts and gene sets, this tool misses very few orthologous genes, recovering practically from single RNA-seq data (Gilbert 2019b). However, some EvidentialGene flaws have been indicated in some experiments. For instance, this program tends to confuse paralogs (genes that do not share coding exon locations) as alternate transcripts (Gilbert 2019a; Gilbert 2019b). This occurs since EvidentialGene classifies loci through CDS alignment. In cases of elevated sequence identity, paralogs align to each other similar to alternate transcripts of one locus (Gilbert 2019a; Gilbert 2019b). Though it is uncommon for RNA-only assemblies, incomplete reconstruction of longer genes has also been indicated by EvidentialGene (Gilbert 2019a). Additionally, a few improvements have been proposed since this software lacks essential parts of gene reconstruction (e.g. verification of errors and contamination or non-coding gene classification), and presents some little developed features, e.g. homology annotation (Gilbert 2019b). Nevertheless, when compared to other methods, EvidentialGene is still more accurate in reconstructing complete transcripts and gene sets and removing redundant sequences.

2.5.1.1.1. EvidentialGene pipeline: tr2aacds

EvidentialGene (or Evigene) usually processes large piles of transcript assemblies through pipelines, particularly the tr2aacds (Gilbert 2016b). This perl script, available at http://arthropods.eugen.es.org/EvidentialGene/about/EvidentialGene_trassembly,

constructs transcripts and genes from biological gene products by taking assemblers (e.g. Trinity, Velvet/O, SOAPdenovo) fasta files as input and generating coding DNA and amino acid sequences through its contigs (Zhang et al. 2019; Chen et al. 2015; Grabherr et al. 2011; Schulz et al. 2012; Luo et al. 2012). In its following methodology steps, tr2aacds requires additional software and programs, namely: 1. *fastanrdb* from to decrease duplicate and similar sequences (redundancy); 2. the previously mentioned cd-hit and cd-hit-est for protein and nucleotide clustering; 3. *blastn* and *makeblastdb* to search regions of local resemblance among sequences (Gilbert 2016b). This process results in transcript outputs being separated into three distinct classes. Main (or Okay) class gathers all best transcripts with the unique coding portions (Zhang et al. 2019). Alternate and Drop classes acquire all possible isoforms and transcripts not passing the internal filter, respectively (Chen et al. 2015). Previous research has demonstrated this pipeline to properly collect transcripts with optimal coding potential to construct optimized transcriptomes and improve transcripts quality and integrity (Chen et al. 2015). Moreover, it has been proven to be highly accurate in redundancy removal and retrieving reliable gene information.

2.5.1.2. Analysis 1.1. Eight *de novo* transcriptome assemblies

As previously mentioned, Trinity assembler outputted a fasta file for each sample, with the reconstructed transcriptomes. During Analysis 1.1, these files were used as input to perform a redundancy assessment and removal. This was conducted through a script (*evigene3.sh*), provided and run through MobaXterm. In this script, EvidentialGene's pipeline tr2aacds was implemented. Figure 20 illustrates the structure of Evigene script, and respective selected arguments and parameters. Sample C10 was used as an example for demonstration.

```
1 module load evigene/20may20
2
3 tr2aacds.pl -mrnaseq /
4     "/mnt/scratch/sbism12/analysis/trinity_analysis1/analysis_1.1/C10/Trinity.fasta" / #input file
5     -MINCDS=50 -NCPUs=16 -MAXMEM=180000 -logfile -tidyup
```

Figure 20 – Redundancy assessment of eight transcriptomes (Analysis 1.1), using tr2aacds pipeline. Overview of the *evigene3.sh* script, and respective parameters specified for the present study. Exemplification using sample C10.

Since only RNA sequences are being considered, *mrnaseq* database was specified. MINCDS parameter allowed to determine the minimum number of base pairs detected in identical coding sequences, posteriorly clustered and removed. For each individual run, four main folders – *okayset*, *dropset*, *inputset* and *tmpfiles* – were generated as output. Within *okayset* folder, a *okay.cds* file containing the unique CDS was used for further evaluation.

2.5.1.3. Analysis 1.2. *de novo* transcriptome assemblies concatenated

During Analysis 1.2, the outputted Trinity's fasta files were combined into a single file, and posteriorly used as input to perform a redundancy assessment and removal. Prior to concatenation, all transcript sequences IDs in each fasta file were differentiated with the addition of distinguished prefixes, through Notepad++. Ensure unique IDs allows to prevent common validation errors, such as duplicates. Similarly to Analysis 1.1, redundancy removal was conducted through *evigene3.sh* script, implementing the *tr2aacds* pipeline. Figure 21 illustrates the structure of Evigene script for this analysis.

```
1 module load evigene/20may20
2
3 tr2aacds.pl -mrnaseq /
4     "/mnt/scratch/sbism12/analysis/trinity_analysis2/reads_concatenated/Trinity.fasta" / #input file
5     -MINCDS=60 -NCPUs=16 -MAXMEM=180000 -logfile -tidyup
```

Figure 21 – Redundancy assessment of eight merged transcriptomes (Analysis 1.2), using *tr2aacds* pipeline. Overview of the *evigene3.sh* script, and respective parameters specified for the present study.

Database and following metrics remained unaltered. Likewise *okayset*, *dropset*, *inputset* and *tmpfiles* folders were outputted. The *okay.cds* file was also used for further evaluation.

2.5.1.4. Analysis 2. Single reconstructed assembly

The process of redundancy evaluation and minimisation conducted in Analysis 2, followed the same methodology presented for Analysis 1.1 and 1.2. Once more, tr2aacds pipeline was implemented through *evigene3.sh* script, and the same output folders and files were generated. Figure 22 illustrates the structure of Evigene script for this analysis.

```
1 module load evigene/20may20
2
3 tr2aacds.pl -mrnaseq /
4             #input file
5             "/mnt/scratch/sbism12/analysis/trinity_analysis1/analysis_1.2/assemblies_concatenated/Trinity.fasta" /
6             -MINCDS=60 -NCPUs=16 -MAXMEM=180000 -logfile -tidyup
```

Figure 22 – Redundancy assessment of a single transcriptome (Analysis 2), using tr2aacds pipeline. Overview of the *evigene3.sh* script, and respective parameters specified for the present study.

2.5.1.5. Metrics to evaluate the overall quality of all Evigene analyses

As indicated above, *okay.cds* files were created as output, at the end of each EvidentialGene analysis. These files were then used as input, to retrieve quality indicator metrics of the Evigene performances in the following command:

```
$ ./contig_stats.pl -f 'okay.cds file' -t 200 -o 'output file'
```

The *-f* and *-t* parameters specify the FASTA file format and the length of the contigs to be counted, respectively. Contigs with a length superior to 200bp were considered.

The evaluating statistics obtained for assessment and comparison are indicated below:

1. maximum contig length;

2. number of contigs > 200;
3. total bases in contigs > 200;
4. N50 for contigs > 200;
5. contigs > 200 in N50;
6. GC contigs > 200;
7. mean length for contigs > 200.

2.5.2. Assessment of Completeness

The term completeness is often described as an estimation of the portion of already found and established genes that are expected to be in a given genome or transcriptome. Currently, many genomic and transcriptomic studies look cataloguing complete repertoires of protein-coding genes, and facilitating completeness assessments in terms of expected gene content (Waterhouse et al. 2019). This prevents incomplete and non-quality data from negatively impacting biological conclusions. In any genome or transcriptome annotation process, it is essential to endue a set of well-characterized genes (Parra, Bradnam, and Korf 2007). However, the fast pace of sequencing and the continuous increase of sequenced genomes and transcriptomes hampers the identification of genes and their uniqueness to a given species or taxonomic group, and complicates genes and resulting products categorization (Parra, Bradnam, and Korf 2007). To outline gene finding issues, multiple computational programs and software have been introduced. These are commonly classified as being alignment-based or composition-based, with others combining both strategies (Parra, Bradnam, and Korf 2007). Briefly, and as mentioned in (Parra, Bradnam, and Korf 2007), methods based on alignments are usually applied to predict protein-coding genes with closed-related homologs. On the other hand, composition-based methods rely exclusively on genes' intrinsic properties (e.g. introns, splice sites, intergenic regions) to construct and predict gene structures. Approaches combining both strategies focus instead on searching a query sequence against a database (with known coding sequences) and then incorporate the matching results into a scoring scheme. Identifying and producing reliable and accurate sets of genes, improves genomes and transcriptomes completeness, as well as the understanding of its encoded genes relations (Parra, Bradnam, and Korf 2007). Orthologs genes are also widely used for gene classification and functional annotation of genomes and transcriptomes. As already mentioned in a previous introductory section, a last common ancestor's single gene gave rise to this homologs genes, present and conserved in many species and lineages (Fitch 1970; Koonin 2005). Tracing orthologs

across a given lineage, allows to infer the core genes of the last common ancestor, as well as its associated protein sequences (Waterhouse, Zdobnov, and Kriventseva 2011). This way, when studying a given organism's genome or transcriptome, its completeness can be evaluated by determining which orthologs genes should be present. Though orthologs relations are not defined by gene function, similar functionality may be inferred through equivalent genes, particularly for single-copy orthologs (Waterhouse, Zdobnov, and Kriventseva 2011). Particularly through species from vertebrate, fungal and arthropod groups, it has been possible to catalog various eukaryotic genes, including orthologs, and identify its functions (Waterhouse, Zdobnov, and Kriventseva 2011). Given the importance of gene content and its association to define complete genomes and transcriptomes, various programs and softwares have been focusing on completeness determination, for instance: CEGMA (presently discontinued), CheckM for prokaryotes, EukCC aimed for microbial eukaryotes, CheckV developed for viruses and BUSCO tool focused on single-copy orthologs (Parra, Bradnam, and Korf 2007; Parks et al. 2015; Saary, Mitchell, and Finn 2020; Nayfach et al. 2021; Manni, Berkeley, Seppey, Simão, et al. 2021).

2.5.2.1. BUSCO: a tool for completeness assessment

The main method currently using established and known genes to implement completeness quantification for assembled genomes and transcriptomes, as well as annotated protein-coding genes, is the Benchmarking Universal Single-Copy Orthologues (BUSCO, <https://busco.ezlab.org/>) tool (Simão et al. 2015). This program mainly focuses on identifying sets of genes only present as single-copy orthologs in a given lineage or taxonomic group (Waterhouse et al. 2019). In other words, it searches for genes that did not suffer loss or duplication events, remaining as single copies in newly sequenced organisms since the last common ancestor. The search for genes under single-copy control is accomplished through a ortholog catalogue designated OrthoDB (Zdobnov et al. 2021). This catalogue, available at <https://www.orthodb.org/>, has highly expanded the number of benchmarking datasets, allowing the cover of major and novel species clades (Manni, Berkeley, Seppey, Simão, et al. 2021). For instance, in the case of arthropods, BUSCO used to provide evaluations for only five lineages: Arthropoda, Insecta, Hymenoptera, Endopterygota and Diptera (Waterhouse et al. 2019). With the most advanced OrthoDB version, the number of lineages was increased to eight (Manni, Berkeley, Seppey, Simão, et al. 2021). More recently, besides large and microbial eukaryotes, and prokaryotes, viral datasets (approximately 27) have been included for assessment (Manni, Berkeley, Seppey, Simão, et al. 2021).

Besides the association with a better catalogue, BUSCO workflow was also improved by implementing the alternative gene predictors: MetaEuk and Prodigal. Briefly and as mentioned in (Levy Karin, Mirdita, and Söding 2020), MetaEuk identifies protein-coding genes with single and multiple exons in eukaryotes. This toolkit is based on a homology approach that permits direct annotations to the discovered genes from the correspondent proteins. MetaEuk has also been focused on collecting proteins from marine eukaryotes on establishing a more comprehensive catalog and database. Prodigal focus on upgrading the identification of existing genes in prokaryotic and microbial genomes, accurately predicting translation initiation sites and reducing the obtainment of false positives (Hyatt et al. 2010). This toolkit coupled with the expanded Bacteria and Archaea datasets, establishes BUSCO as an optimal choice to assess prokaryotic data (Manni, Berkeley, Seppey, Simão, et al. 2021).

As mentioned above, BUSCO tool offers different workflows, dependent on the input data for analysis: genome, gene set or transcriptome. Figure 23 represents this tool assessment process, with correspondent relative run times.

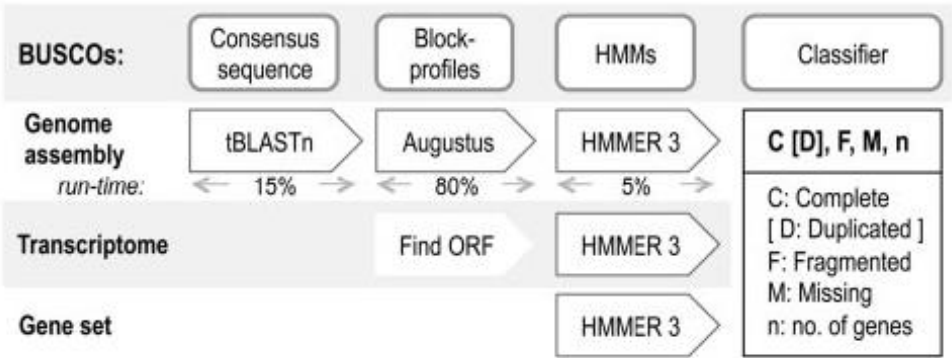


Figure 23 – BUSCO assessment workflows (genome, transcriptome and gene set), and associated third-party components. Image retrieved from (Simão et al. 2015).

Common to all workflows is the implementation of profile hidden Markov models, or profile HMMs (Eddy 2011). This probabilistic and inference approach classifies and assesses whether BUSCO gene matches are or not orthologs based on defined score and length cut-offs (Simão et al. 2015). An advantageous aspect of HMM algorithm is the computation and sum of various alignment probabilities instead of a unique best-scoring alignment (Eddy 2011). This reduces potential identification bias and captures the conserved protein sequences across all sets (Eddy 2011; Seppey, Manni, and

Zdobnov 2019). When matches meet BUSCO HMM score cut-offs, four main classifications can be obtained: 1. Complete, subdivided into 'Complete and single-copy' and 'Complete and duplicated'; 2. Fragmented; 3. Missing (Manni, Berkeley, Seppey, and Zdobnov 2021). Complete classification is attributed to genes whose total lengths fall within BUSCO profile expectations (Waterhouse et al. 2019). The differentiation between single-copy and duplicated is based on the number of correspondent appearances. Single-copy refers to complete matches appearing only once. Duplicated indicates complete matches appearing more than one time. Fragmented classification is attributed to partial matches. Partiality is considered when more than 10% of the gene length does not meet BUSCO profile length expectations (Jauhal and Newcomb 2021). Lastly, matches not passing the thresholds are labelled as missing (Jauhal and Newcomb 2021). In the case of transcriptomes, the simple longest open reading frame (ORF) technique is also applied for a prior and initial gene annotation (Simão et al. 2015). For contextualization, ORFs are regions of DNA located between a start and a stop codon, since they do not contain stop codons when translated into amino acids and proteins (Sieber, Platzer, and Schuster 2018). The longest open reading frame in eukaryotes is usually used for gene translation (Sieber, Platzer, and Schuster 2018). When assessing a genome, gene annotation is primarily conducted using Augustus predictor and block-profiles (Simão et al. 2015; Keller et al. 2011; Stanke and Morgenstern 2005). The term block is allusive to ungapped and highly conserved regions of multiple sequence alignments (Keller et al. 2011). A block profile is a collection of each aminoacid specific position and distribution in a block (Keller et al. 2011). Augustus employs block profiles, as protein signatures, in potential matches identified using tBLASTn (Keller et al. 2011; Simão et al. 2015). This improves gene prediction accuracy in eukaryotes (Waterhouse et al. 2019). For contextualization, tBLASTn is a tool that uses a protein query to identify similar database encoding-protein sequences (Gertz et al. 2006).

As with any other method, BUSCO also presents a few biases and limitations. For instance, (Simão et al. 2015) pointed out some drawbacks at BUSCO's gene prediction step, when genome or transcriptome assemblies appear less complete than their corresponding genes. Moreover, it indicated that the inverse might suggest missing BUSCO gene matches on the annotated gene set that are, in fact, present in the genome or transcriptome. Another example is the experimental research conducted in (Waterhouse et al. 2018), which demonstrated BUSCO assessments still falling short of 100% correct classification, since accomplishing gene prediction and orthology attribution remains very challenging. To support this statement, this study pointed that some BUSCOs could be incorrectly classified as missing since they could be to

divergent, or present complex gene structures hampering its location and prediction. Additionally, if the assembly process fails to disintegrate heterozygous alleles, BUSCOs can also be inaccurately classified as duplicates. Likewise, (Jauhal and Newcomb 2021) listed some inconveniences of this tool. For instance, a gene that is fragmented may be unduly labelled as missing if it does not meet the partial match threshold. Also affecting BUSCO scores and quality is the naturally defective performance of some taxa when compared to others. To determine whether a score is satisfactory or not, this study recommends the inclusion of related species assemblies scores for comparison, as some organisms groups may perform better than others. Though some improvements and disadvantages were proposed and outlined for BUSCO, this tool still provides fair comparisons and high consistency across different species (Simão et al. 2015). It also remains suited for accurate and efficient completeness assessments of genomes, transcriptomes or gene sets in both eukaryotic and prokaryotic species (and, more recently, viruses).

2.5.2.1.1. Setup and implementation of BUSCO software

BUSCO is available and can be downloaded from a public GitLab platform. Its source code is also maintained through distribution channels on Bioconda (<https://anaconda.org/bioconda/busco>), and as a Docker container (Manni, Berkeley, Seppey, Simão, et al. 2021). Third-party components (e.g. HMM, Augustus, Prodigal, MetaEuk) must be added according to the sequences under analysis (Eddy 2011; Stanke and Morgenstern 2005; Levy Karin, Mirdita, and Söding 2020; Hyatt et al. 2010). MobaXterm terminal is another alternative to run BUSCO, since it provides an engine container to Docker, known as 'CharlieCloud' (Priedhorsky and Randles 2017). This container (<https://wikis.cf.ac.uk/confluence/display/BW/Busco+-+Evaluation+of+assemblies>) is suited for high-performance computing (HPC) environments and provides unprivileged permissions, using user namespaces in the Linux kernel.

For genome, transcriptome or gene set assessments, BUSCO requires input sequence data to be provided in the standard FASTA format (Manni, Berkeley, Seppey, and Zdobnov 2021). Moreover, it also requires the definition and providence of four mandatory arguments (Manni, Berkeley, Seppey, and Zdobnov 2021), namely:

1. *-i* (or *-in*), to define the input fasta file for analysis;

2. *-o* (or *- -out*), to specify the output directory folder, holding intermediate data and all generated results;
3. *-l* (or *- -lineage*), to determine the BUSCO lineage for assessment;
4. *-m* (or *- -mode*), to set the assessment mode: genome, transcriptome or proteins.

As previously mentioned, BUSCO results are simplified and outputted into the categories 'complete and single-copy', 'complete and duplicated', 'fragmented' and 'missing'. This are reported in percentages, and can be summarized in a graphic.

2.5.2.2. Completeness assessment using BUSCO tool

As previously mentioned, EvidentialGene software outputted *okay.cds* files, containing the best set of putative assembled transcripts in each conducted analysis. These files were used as input to perform completeness assessments using BUSCO tool. BUSCO was launched through 'CharlieCloud' container, through MobaXterm terminal. To launch and invoke this container, two commands were used:

```
$ /mnt/scratch/nodelete/CharlieCloud/scripts/CharlieCloud_busco.interactive.job
```

```
$ srun -c 8 --mem 16G --pty  
/mnt/scratch/nodelete/CharlieCloud/scripts/CharlieCloud_busco.interactive.job
```

Once inside 'CharlieCloud' container, all best sets of transcripts with unique codifying regions were assessed through the default command line - "*busco -i <sequence file> -l <lineage> -o <output_name> -m <mode> <other options>*" - yet its arguments were adapted and specified to the present thesis as followed:

1. *-i* (or *- -in*) remained unaltered. Through this argument, the *okay.cds* files were specified as input;
2. *-l* (or *- -lineage*) was altered to *--auto-lineage-euk*, since this speciation only considers the eukaryote tree to find optimum lineage paths;
3. *-o* (or *- -out*) remained unaltered. This argument determined the path to the generated output directories;

4. *-m* (or *-mode*), argument that defines the BUSCO analysis mode to run, was specified to *tran* (or *transcriptome*), since transcriptome assemblies are being considered.

Thus, the BUSCO assessment command line was:

```
$ busco -i <okay.cds fasta file> - -auto-lineage-euk -o <output folder> -m tran
```

At the end of each run and analysis, a final report was provided. Results were classified as 'complete and single-copy', 'complete and duplicated', 'fragmented' and 'missing'. These were reported in percentages.

2.5.2.3. Pre-Evigene metrics for an overall quality comparison

EvidentialGene statistics command was also applied in the fasta files outputted at the end of Analyses 1 and 2, using Trinity assembler. This was performed to compare the results obtained for reconstructed transcriptomes solely assembled (pre-evigene), and transcripts passed through the process of redundancy removal (post-evigene). With this comparison, it was expected to understand the effects of transcriptome selection and evaluate the quality tendency. The command to retrieve Evigene metrics described above was altered to:

```
$ ./contig_stats.pl -f 'Trinity fasta file' -t 200 -o 'output file'
```

Completeness assessments using BUSCO tool were also performed in the fasta files produced by Trinity at the end of Analyses 1 and 2. 'CharlieCloud' container was equally launched and invoked. Once inside the container, the following command line was used:

```
$ busco -i <Trinity fasta file> - -auto-lineage-euk -o <output folder> -m tran
```

The results were also reported in four main categories and summarised for comparison.

Chapter III. Results and Discussion

1. Sequencing quality assessment

FastQC summarised an evaluation for forward and reverse raw reads of each sample for the nine available modules. These are illustrated in Table 1. Based on the quality level of the sequenced contigs, a module would raise one of three available categories, namely:

1. a 'Pass' for good quality sequences – highlighted in green in Table 1;
2. a 'Warning' for sequences with reasonable quality – highlighted in orange in Table 1;
3. a 'Fail' for poor quality sequences – highlighted in red in Table 1.

FASTQC

Name: Evaluation of samples collected from *Carcinus maenas*

Overall Results

Name	Per Base Sequence Quality	Per Sequence Quality Scores	Per Base Sequence Content	Per Sequence GC Content	Per Base N Content
C10_forward_reads	PASS ●	PASS ●	FAIL ●	WARNING ●	PASS ●
C10_reverse_reads	PASS ●	PASS ●	FAIL ●	WARNING ●	PASS ●
C13_forward_reads	PASS ●	PASS ●	FAIL ●	FAIL ●	PASS ●
C13_reverse_reads	PASS ●	PASS ●	FAIL ●	FAIL ●	PASS ●
C3_forward_reads	PASS ●	PASS ●	FAIL ●	FAIL ●	PASS ●
C3_reverse_reads	PASS ●	PASS ●	FAIL ●	FAIL ●	PASS ●

G1R1_forward_reads	PASS ●	PASS ●	FAIL ●	FAIL ●	PASS ●	
G1R1_reverse_reads	FAIL ●	PASS ●	FAIL ●	FAIL ●	PASS ●	
R3_forward_reads	PASS ●	PASS ●	FAIL ●	FAIL ●	PASS ●	
R3_reverse_reads	PASS ●	PASS ●	FAIL ●	FAIL ●	PASS ●	
R11_forward_reads	PASS ●	PASS ●	FAIL ●	WARNING ●	PASS ●	
R11_reverse_reads	PASS ●	PASS ●	FAIL ●	WARNING ●	PASS ●	
R12_forward_reads	PASS ●	PASS ●	FAIL ●	FAIL ●	PASS ●	
R12_reverse_reads	PASS ●	PASS ●	FAIL ●	FAIL ●	PASS ●	
R15_forward_reads	PASS ●	PASS ●	FAIL ●	FAIL ●	PASS ●	
R15_reverse_reads	PASS ●	PASS ●	FAIL ●	FAIL ●	PASS ●	
Name	Sequence Length Distribution	Adapter Content	Overrepresented Sequences	Sequence Duplication Levels		
	C10_forward_reads	PASS ●	PASS ●	FAIL ●		FAIL ●
	C10_reverse_reads	PASS ●	PASS ●	FAIL ●		FAIL ●
	C13_forward_reads	PASS ●	PASS ●	WARNING ●		FAIL ●
	C13_reverse_reads	PASS ●	PASS ●	WARNING ●		FAIL ●
	C3_forward_reads	PASS ●	PASS ●	FAIL ●		FAIL ●
	C3_reverse_reads	PASS ●	PASS ●	FAIL ●		FAIL ●

G1R1_forward_reads	PASS ●	PASS ●	WARNING ●	FAIL ●
G1R1_reverse_reads	PASS ●	PASS ●	WARNING ●	FAIL ●
R3_forward_reads	PASS ●	PASS ●	FAIL ●	FAIL ●
R3_reverse_reads	PASS ●	PASS ●	FAIL ●	FAIL ●
R11_forward_reads	PASS ●	PASS ●	FAIL ●	FAIL ●
R11_reverse_reads	PASS ●	PASS ●	FAIL ●	FAIL ●
R12_forward_reads	PASS ●	PASS ●	WARNING ●	FAIL ●
R12_reverse_reads	PASS ●	PASS ●	WARNING ●	FAIL ●
R15_forward_reads	PASS ●	PASS ●	FAIL ●	FAIL ●
R15_reverse_reads	PASS ●	PASS ●	FAIL ●	FAIL ●

Table 1 – Summarized evaluation of each samples' forward and reverse reads, for the nine available modules, using FastQC.

Four modules assessed all samples' contigs as good quality sequences, in particular: 1. 'Per Sequence Quality Scores'; 2. 'Per Base N Content'; 3. 'Sequence Length Distribution'; 4. 'Adapter Content'. Moreover, they implied a diminished percentage of N base calls per position. A successful adapters removal for each sample's forward and reverse reads was confirmed with the positive results obtained for the 'Adapter Content' module.

As referred to in (FastQC 2020), the 'Per Base Sequence Quality' module is commonly evaluated through a BoxWhisker plot. Quality scores are represented on the y-axis, and each base position in a read on the x-axis. This diagram divides the y-axis into three main portions, namely: good, reasonable and poor. According to a quality

call, data displays in one of these categories. Additionally, median values, second and third quartile percentages, and the interquartile range are calculated for each base position on a read. Since quality tends to degrade over the duration of sequencing runs, it is common to observe bases positioned at the end of a sequence falling into the reasonable and poor section of the plot. In the present work, this module solely failed the reverse reads of sample G1R1, suggesting lower quality of its bases, and problems in the library construction.

The remaining modules classified all sample sequences as poor or reasonable quality reads. As referred to in (FastQC 2020), 'Per Base Sequence Content' module plots the amount of A, T, G and C base pairs within a sequence. In an unbiased library, it is expected to observe little to no difference among bases. The amount of base pairs is often represented as parallel lines in the plot. In the present work, this module failed all samples forward and reverse reads. This result suggested the presence of considerable differences (probably higher than 20%) between bases A and T, or G and C, in any read position. In the case of invertebrates, such as the crustaceans, AT bias are expected. 'Sequence Duplication Levels' module also failed all reads under analysis. Its graphical representation allows scoring the relative number of sequences with distinct levels of duplication (FastQC 2020). This result suggested an elevated number of reads presenting duplicates and, consequently, a low degree of coverage of the reference sequences. Library duplicates may appear due to problems during PCR amplification (e.g. artefacts) or biological causes, e.g. random selection of different copies of the same sequence (FastQC 2020). 'Per Sequence GC Content' module raised a 'warning' to both forward and reverse reads of the samples C10 and R11. All sequences of the remaining samples were classified as poor quality contigs. This module plot measures the quantity of GC across the total length of a sequence and compares it against a modelled normal distribution (FastQC 2020). In the present work, the non-positive results obtained for this module reinforced a possible biased and poor-quality library construction. Lastly, the 'Overrepresented Sequences' module classified three samples (C13, G1R1, R12) contigs with reasonable quality. All remaining samples forward and reverse reads raised a 'fail'. As referred to in (FastQC 2020), these classifications are commonly attributed when more than 0.1% and 1% of the total sequences demonstrate overrepresentation, respectively. Overrepresented reads are frequently indicative, once more, of contamination during the library preparation, and little diversity among the set.

It is important to consider that this module's results, though not following the expected standard or established normality, may be of interest depending on the study in question and respective purpose for analysis. In the case of the present work, the

predominance of non-positive results allowed to assume the occurrence of frequent errors during the preparation and construction of the initial library, as well as persistent problems during sequencing.

Graphical representations for the 'Per Base Sequence Content', 'Per Sequence GC Content' and 'Sequence Duplication Levels' modules were obtained for the contaminated samples C10 and C13. These can be consulted on the Appendix Supplementary Figures 1 to 6.

2. Transcriptome sequencing output and Trimming

The high-throughput sequencing produced a total of 246,539,629 reads. Table 2 indicates the initial number of input read pairs for each collected reference and contaminated sample. It was expected to observe an approximate number of sequenced reads for all samples. The majority of samples presented proximately 31 million read pairs. Nonetheless, samples C10 and G1R1 exhibited a lower number of reads, approximately 25 and 18 million read pairs, representing a difference of near 19% and 42%, respectively. On the contrary, the sample R15 presented an increase of approximately 48% of initial reads. This discrepancy may have occurred due to errors during the sequencing process since 1. crabs collection was conducted under identical conditions, 2. RNA isolation and extraction followed the same methodology for all samples, and 3. the assessment of purity and integrity suggested good quality of all extracted RNA. Regarding the implementation of trimming, there was a slight reduction of reads from the original input set for all extracted samples. The percentage of eliminated reads is represented in the columns with forward and reverse survived reads, and dropped reads. Each sample preserved above 93% of both forward and reverse reads. These were considered in the following methodology steps. Though the inclusion of the trimming step enabled the elimination of a significant percentage of lower-quality reads, this was not sufficient to improve the overall quality of the set.

<i>Trimmomatic and Trimming Analysis</i>						
<i>Moderately Contaminated or Reference Estuary</i>	<i>Sample</i>	<i>Input Read Pairs</i>	<i>Both Surviving</i>	<i>Forward Only Surviving</i>	<i>Reverse Only Surviving</i>	<i>Dropped</i>
<i>Contaminated</i>	<i>C10</i>	24873597	24018542 (96.56%)	645291 (2.59%)	157653 (0.63%)	52111 (0.21%)
<i>Contaminated</i>	<i>C13</i>	31468866	29608189 (94.09%)	1536263 (4.88%)	185459 (0.59%)	138955 (0.44%)
<i>Contaminated</i>	<i>C3</i>	31292616	30245758 (96.65%)	827440 (2.64%)	157051 (0.50%)	62367 (0.20%)
<i>Contaminated</i>	<i>G1R1</i>	18394577	17503498 (95.16%)	721693 (3.92%)	109421 (0.59%)	59965 (0.33%)
<i>Reference</i>	<i>R11</i>	31639689	30222732 (95.52%)	1062863 (3.36%)	275309 (0.87%)	78785 (0.25%)
<i>Reference</i>	<i>R12</i>	30253825	28514299 (94.25%)	1424382 (4.71%)	189620 (0.63%)	125524 (0.41%)
<i>Reference</i>	<i>R15</i>	46122392	43258630 (93.79%)	2211189 (4.79%)	458388 (0.99%)	194185 (0.42%)
<i>Reference</i>	<i>R3</i>	32494067	31344612 (96.46%)	889462 (2.74%)	196711 (0.61%)	63282 (0.19%)

Table 2 – Transcriptome Sequencing and Trimming results per sample.

3. *de novo* Transcriptome Assemblies output

The statistics provided from the supplementary script *TrinityStats.sh*, to evaluate the overall quality of the reconstructed *de novo* assemblies during Analyses 1 and 2 from Workflow II, are described below. Figures 24 and 25 illustrate the total number of assembled reads and total assembled transcripts per sample, respectively. Table 3 counts the contig average and median length per assembly, the percentage of GC and metrics N10 to N50 for contiguity assessment.

Concerning Figure 24, eight individual transcriptomes were assembled during Analysis 1. The total number of reads per sample decreased compared to the initial number of input read pairs posteriorly submitted to trimming. Additionally to this significant reduction, a considerable discrepancy among the counted number of reads is likewise quite evident: 1. reference samples R11 and R3 present less than 15M reads; 2. samples R12 and R15 count 40M to 70M reads; 3. sample C3 presents less than 10M reads; 4. samples C10 and C13 count 40M to 80M reads; 5. and sample G1R1 presents an approximate value of 120M of reads. Samples G1R1 and C13 exhibit the highest number of reads, suggesting that they were better sequenced. This contrast is still perceptible when comparing the two groups of samples – reference and contaminated. Despite being a primary and quite superficial indicator, the observed discrepancy in sequencing depth suggests a low quality of the assemblies. The intended result would be a uniform distribution, regardless of high or low values. Regarding Analysis 2, only a single transcriptome was assembled since all samples' reads were concatenated (Figure 24). A higher count of total reads was observed compared to each assembly. The objective of performing a second analysis was to verify if merging all raw reads would lead to a better *de novo* assembly. The concatenation process may or may not lead to a better-reconstructed transcriptome to use as a reference (Chessa 2021). In the present work, merging raw reads appears to have improved the final assembled transcript.

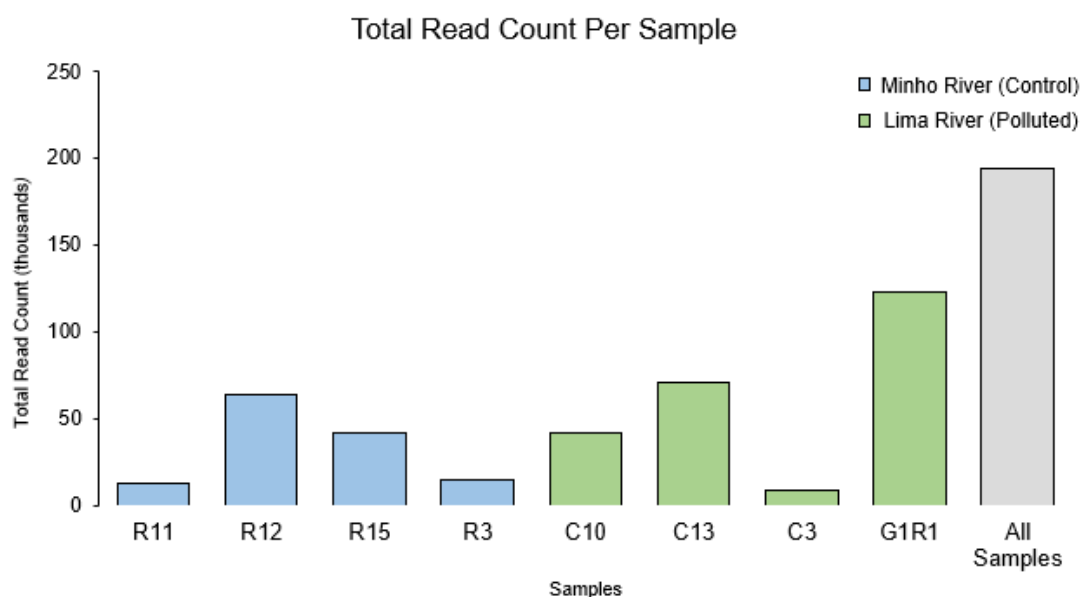


Figure 24 – Graphical representation of the total assembled reads per reconstructed assembly (Analyses 1 and 2).

In relation to Figure 25, the total assembled transcripts count followed a similar distribution to Figure 24 (Analyses 1 and 2). The discrepancy among samples of the same set and groups is very apparent, with sample G1R1 presenting over 90 million assembled transcripts and C3 sample presenting the lowest count, less than 5 million assembled transcripts. Samples G1R1 and C13 exhibit the highest number of assembled transcripts, suggesting that they were better sequenced. Since it is notorious that sample G1R1 highly contributes to a non-regular distribution and a high discrepancy among samples, its removal in further analysis could allow to retrieve more reliable biological conclusions and results. Regarding Analysis 2, a quite superior number of reconstructed transcripts (180 million) was counted when compared to each individual assembly (Figure 25). Once again, the process of concatenation appears to have improved the final *de novo* assembly.

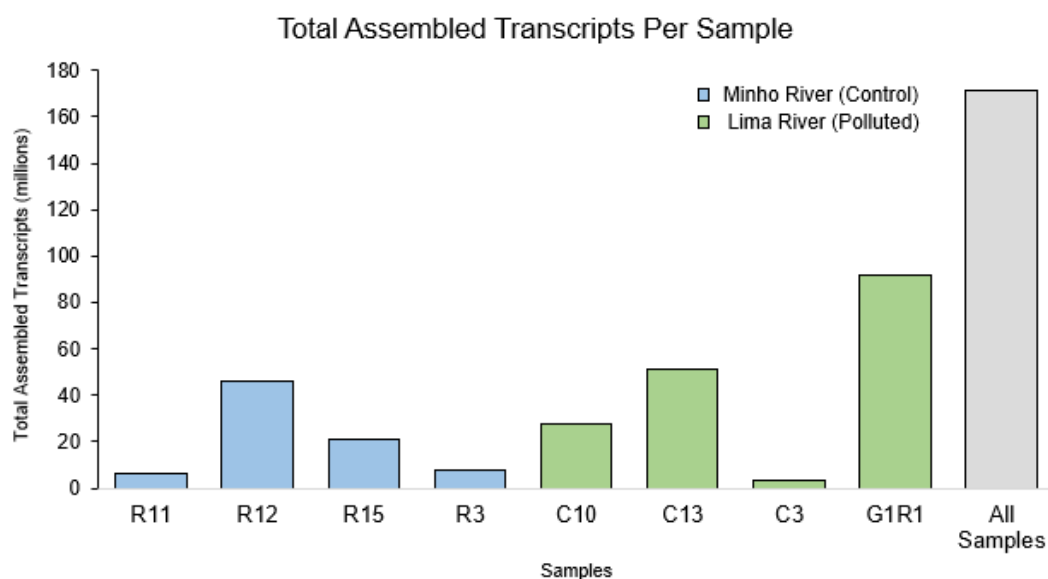


Figure 25 – Graphical representation of the total assembled transcripts per reconstructed assembly (Analyses 1 and 2).

In relation to Table 3 and Analysis 1, a discrepancy is still very apparent. Sample C13 exhibits the highest value for the N50 statistic, suggesting that at least half of its contigs present the highest length compared to all remaining individual samples under analysis. Sample G1R1 follows with a N50 value of 1130 bp. The lowest N50 count is associated with sample C3 (539 bp). As previously indicated, this metric was extended to provide the N10 to N40 thresholds with similar meanings. In the case of the N10 statistic, only 10% of the total contigs are being considered. Samples C13 and R12 exhibited the highest values, counting 3440 bp and 3217 bp, respectively. The lowest count of this threshold is associated to the reference sample R11, which is, 1165 bp. In the case of the median contig length and average contig statistics, the highest and lowest values are associated to samples G1R1 and C3, respectively. The percentage of GC was presented exclusively as an informative metric.

Regarding the single reconstructed assembly from Analysis 2, referred to as '*Merged Samples*', the highest values were obtained for the N10 to N50 thresholds, as well as for the median contig length and average contig statistics, when compared to each individual assembly. The N50 statistic indicated that at least half of this assembly contigs have the highest length (1464 bp). These results suggest that when all raw reads are firstly merged, it is more likely to obtain a more complete and better

reconstructed transcriptome. These results are in conformity with the inferences extracted from Figures 24 and 25.

<i>de novo Transcriptome Assembly using Trinity®</i>									
<i>Moderately Contaminated or Reference Estuary</i>	<i>Sample</i>	<i>Stats Based On All Transcript Contigs</i>							
		<i>Contig N10</i>	<i>Contig N20</i>	<i>Contig N30</i>	<i>Contig N40</i>	<i>Contig N50</i>	<i>Median Contig Length</i>	<i>Average Contig</i>	<i>GC %</i>
<i>Contaminated</i>	<i>C10</i>	2565	1936	1532	1229	963	441	665.64	41.75%
<i>Contaminated</i>	<i>C13</i>	3440	2530	1950	1514	1160	437	732.66	42.64%
<i>Contaminated</i>	<i>C3</i>	1369	1005	776	641	539	345	456.74	42.64%
<i>Contaminated</i>	<i>G1R1</i>	2777	2093	1695	1391	1130	490	745.23	42.46%
<i>Reference</i>	<i>R11</i>	1165	913	763	653	566	416	486.31	39.97%
<i>Reference</i>	<i>R12</i>	3217	2371	1853	1441	1113	434	715.35	42.66%
<i>Reference</i>	<i>R15</i>	1503	1141	932	778	644	389	516.64	41.42%
<i>Reference</i>	<i>R3</i>	1708	1246	994	804	664	403	532.55	40.51%
<i>Both Estuaries</i>	<i>Merged Samples</i>	4303	3010	2301	1841	1464	509	881.16	42.65%

Table 3 - Trinity quality evaluator metrics. Results obtained per reconstructed transcriptome assembly (Analyses 1 and 2). Statistics based on all transcript contigs.

3.1. PCA, MA Plot, Dendrogram and Heatmap

Principal component analysis (PCA) plot was obtained using, as variables, each assembled transcriptome and its gene variation within each sample. The PCA plot is illustrated in Figure 26. This analysis evidenced a complete isolation of the sample G1R1, supporting its previous observed contribution to a high discrepancy among samples. The remaining samples clustered in three small groups, each constituted with at least a reference and a contaminant sample:

1. cluster 1 – reference sample R12 and contaminant sample C13;
2. cluster 2 – reference sample R15 and contaminant sample C10;
3. cluster 3 – reference samples R3 and R11, and contaminant sample C3.

The conjunction of samples of each computed cluster follows and resembles the distribution of Figures 24 and 25. For instance, samples R12 and C13 exhibited very similar counts of total reads and assembled transcripts. Simultaneously, its distribution along the principal components PC1 and PC2 was nearly identical, leading to the recognition of a cluster. The same occurred in clusters 2 and 3. The formation of these agglomerates, and the absence of a uniform distribution to either side of the 0,0 coordinate, may suggest that: 1. crabs individuals are from very close populations; 2. *Carcinus maenas* populations are starting to occupy more of the collection site or presenting a higher percentage in that site/area; 3. samples are very similar, regardless of whether they were collected from an estuary under moderate contamination or an estuary under lower levels of human intervention. Through this analysis, it was expected to identify clusters separating samples from the Minho estuary and samples from the Lima estuary. Instead, their high disparity was observed with the conjunction of samples from both estuaries, suggestion low quality of the sequences. The visible absence of coherence among all samples in the PCA plot is also in accordance with the discrepancy illustrated in Figures 24 and 25, as well as in Table 3.

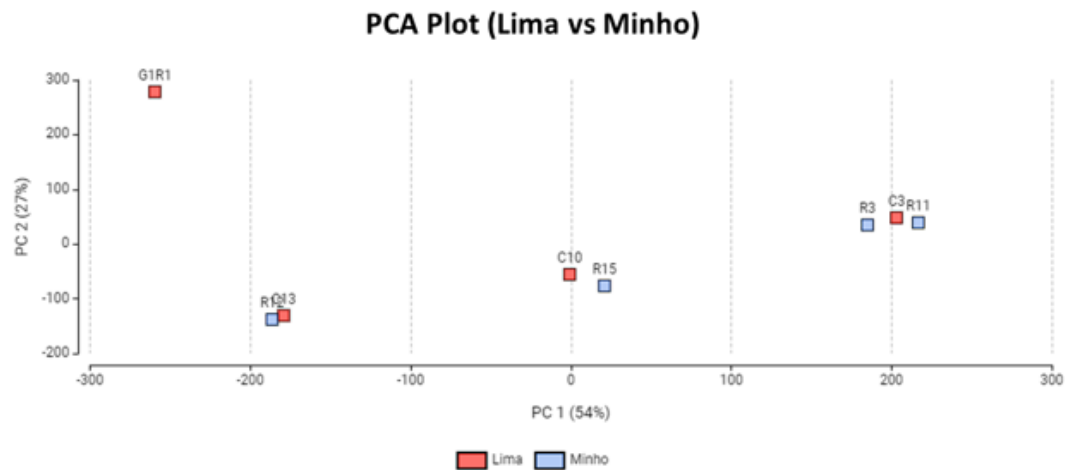


Figure 26 – Visualization of two-dimensional PCA plot. Samples are distributed along PC1 and PC2 (first and second principal component, respectively) axis.

The MA plot, illustrated in Figure 27, evidence a decreased horizontal distribution of the majority of data points along the x-axis. Most values highly accumulate near the y-axis 0,0 coordinate, suggesting that the two treatment groups (Minho and Lima estuaries) appear to be similar in expression patterns. Likewise, the generality of data fall evenly between the established -1 and 1 thresholds on the y-axis. Data points outside the established threshold represent genes with highly differential expression levels. Values falling above the one threshold on the y-axis indicate the number of genes transcripts being more expressed (upregulated) in the Lima river when compared to the Minho river. Oppositely, values below -1 indicate the number of downregulated transcripts. Since the False Discovery Rate (FDR) was established to values below 0.05, not all significantly different genes, appear beyond the 1 and -1 thresholds. Those that appear are represented as coloured points in the plot. These differential genes do not follow most data distribution and are far beyond the y-axis settled threshold. Hence, due to the high variability among the data, the differences between the samples from both estuaries is minimal, not being possible to make a correct analysis.

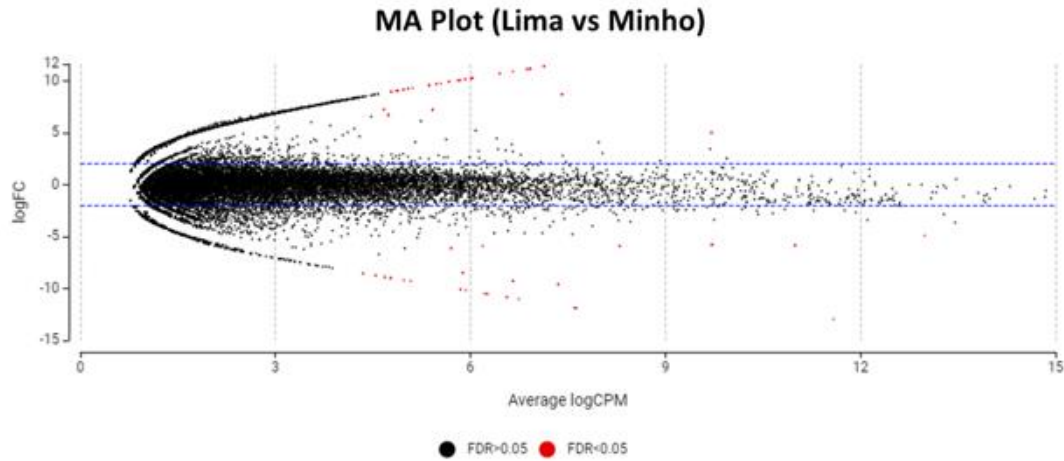


Figure 27 – Visualization of two-dimensional MA plot. Samples are distributed along *logFC* and *Average logCPM* axis. Lines coloured in blue represent the established -1 and 1 thresholds. Points coloured in red represent genes transcripts with below-threshold adjusted p-values.

The dendrogram and heatmap, illustrated in Figure 28, represents the Lima river and the Minho river clusters, as well as the up- and down-regulated genes from the crabs collected from its estuaries. Up-regulated genes (highlighted in red) correspond to genes that are more expressed in organisms collected from the Lima estuary when compared with the Minho estuary. Down-regulated genes (highlighted in green) correspond to genes that are less expressed in organisms collected in the Lima estuary when compared with the Minho estuary. Both clusters were expected to be entirely separate. Samples collected from the Lima river should be separated from the samples collected from the Minho river. However, this was not observed since sample C13 was grouped with three reference samples (R3, R12 and R15), and reference sample R11 was grouped with the remaining contaminant samples (C3, C10 and G1R1). Hereafter, these groups will be referred to as cluster 1 and cluster 2, respectively. This unexpected distribution of the samples C13 and R11 may be explained by the high variability of the sequences described previously. Nonetheless since the organisms travel within the river may have been exposed to a longer or shorter period to the stressors described in each river, thus showing some related patterns between both rivers.

Regarding the up- and down-regulated genes, it is possible to observe the existence of genes with distinct expression levels along the heatmap, given its collection from crabs exposed to low or moderate contamination. As mentioned above, samples C13 and R11 were an exception to this differential transcriptomic expression. The extracted

samples from cluster 1 present a predominance of up-regulated genes. Contrarily, samples from cluster 2 present a predominance of down-regulated genes. Though it is not possible to establish which genes are being repressed or highly expressed, previous ecotoxicological studies have identified and suggested the main metabolic pathways and respective regulators being impacted in the green crab *Carcinus maenas*, when exposed to mixtures of contaminants. Several biomarkers related to neurotransmission, biotransformation, energy metabolism, antioxidant defences and oxidative stress pathways have been identified as early warning indicators of exposure to contamination on this species (Mesquita, Guilhermino, and Guimarães 2011; Rodrigues et al. 2015). For instance, acetylcholinesterase (AChE), an enzyme participating in cholinergic transmission by hydrolysing the neurotransmitter acetylcholine into choline and acetic acid, was shown to be very useful in assessing neurotoxicity (Xuereb et al. 2009). Alterations in AChE activity may result in crabs' behavioural abnormalities such as impairment in locomotion and swimming, with consequent increase of vulnerability to predators (Rodrigues et al. 2015; Simmers and Bush 1983; Sorenson 1973; Vieira et al. 2009), and uncommon reproductive behaviours (Sorenson 1973; Mesquita, Guilhermino, and Guimarães 2011; Rodrigues, Gravato, and Guimaraes 2013). Another example are the Glutathione-S-transferases (GST), catalyst enzymes involved in biotransformation reactions. These have an important role in xenobiotics detoxification and protection against oxidative damage (Rodrigues et al. 2015). Upon pollutants exposure, GST catalyses conjugation of glutathione (GSH), a non-enzymatic defensor against reactive oxygen species, into its oxidised form GSSG (Rodrigues et al. 2015; Forman, Zhang, and Rinna 2009). Enzyme N-acetyl-b-glucosaminidase (NAGase), although not so used, is also significant since it plays a role in the moulting process (Mesquita, Guilhermino, and Guimarães 2011). NAGase, or chitinase, is a chitinolytic enzyme mostly found in crustacean epidermis, responsible to degrade chitin oligosaccharides into monomers that will be re-absorbed during the formation of the new cuticle (Espie and Roff 1995; Meng and Zou 2009). Several studies demonstrated that shore crabs exposed to disrupting compounds and contaminants presented an inhibition of their activity (Meng and Zou 2009). Other biomarkers related to energy metabolism, antioxidant defences and oxidative stress responses are the activity of various other enzymes, such as, lactate dehydrogenase (LDH) and NADP⁺-dependent isocitrate dehydrogenase (ICDHs), catalase (CAT), total glutathione (TG) and superoxide dismutase or SOD (Rodrigues and Pardal 2014; Rodrigues et al. 2015). Additionally, *Carcinus maenas* responses to pollutants may be evaluated through the estimation of specific proteins such as metallothioneins, or MTs (Rodrigues and Pardal 2014). Metallothioneins

represent the principal protective mechanism of a cell against the action of free ions of toxic metals (Amiard et al. 2006). Previous research experiments found correlations between MTs and soluble essential (Cu and Zn) and nonessential (cadmium, Cd) metal concentrations in this species, confirming metal regulation by these proteins (Rodrigues and Pardal 2014; Dziegiel 2004).

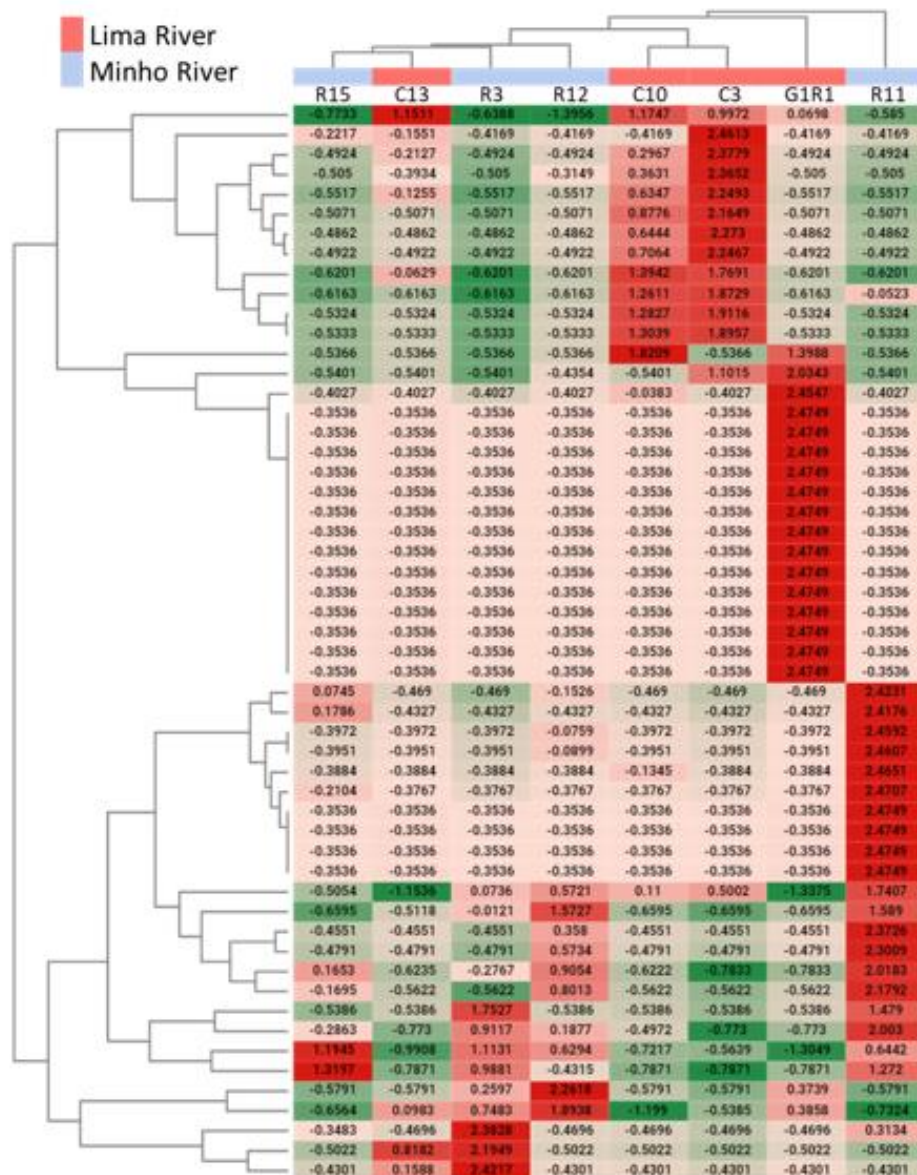


Figure 28 – Dendrogram and heatmap of transcripts of the Lima river compared to the transcripts of the Minho river. Identification of up-regulated (coloured in red) and down-regulated genes (coloured in green) per sample.

4. Assemblies redundancy output (and pre-evigene comparison)

As described in the Materials and Methods section, three separate analyses for removing redundancy were conducted using EvidentialGene software, depending on the merge or non-concatenation of the initial raw reads or the assembled transcripts. Analysis 1 was partitioned since the assessment of redundancy was conducted upon each individual reconstructed transcriptome (Analysis 1.1.), and upon all assemblies concatenated into a single file (Analysis 1.2.). The third analysis was performed on the single transcriptome assembled during Analysis 2. Redundant sequences were eliminated by implementing the tr2aaccs pipeline. Posteriorly, statistics to assess each Evigene performance were obtained through a *contig_stats* command. In Table 4, *pos_evigene* refers to the post-evigene assessment data obtained for all three analyses. In order to observe the effects of minimising redundancy among the sets of contigs, it was necessary to establish a comparison between transcripts submitted to the process of redundancy removal (post-evigene), and transcripts solely assembled (pre-evigene). To achieve this, the *contig_stats* command was once more used, however using the fasta files outputted by Trinity as input. The pre-evigene assessment data obtained for all three analyses is referred to as *pre_evigene* in Table 4.

Regarding Analysis 1.1, there is an evident disproportion in the number of initial transcripts for all samples. These disparity is in accordance with the inferences extracted from the previous results. The sequencing errors led to disparity and low-quality among the sequenced RNA. A significant reduction in the number of final transcripts was observed for all samples under analysis when comparing the pre- and post-evigene results. As it can be observed through the 'Num contigs > 200' statistic, the sample G1R1 exhibited a higher reduction of duplicated transcripts (67.8%). The sample C13 and the reference sample R12 followed with a reduction of proximately 46.2% and 40.7%, respectively. Contrarily, sample C3 exhibited the lowest percentage of eliminated sequences, proximately 5.6%. In relation to the 'Max contig length' statistic, and once concluded the removal of redundancy, it can be observed that the samples G1R1 and C13 present the longest contigs (14502 base pairs). The opposite occurs to sample C3, with its contigs exhibiting a medium length of 3192 base pairs. Through the analysis of both these metrics, it was possible to verify that the EvidentialGene software, and its CDS approach, successfully removed similar and duplicated sequences from the sets of all samples. The remaining metrics were presented exclusively as informative statistics. Moreover, the high percentages of eliminated reads observed for the majority of samples support the predominance of

discrepancy and low-quality of the RNA. When assembling a transcriptome, it is common to observe redundancy and duplicates among reads since transcripts are commonly over-sequenced. This analysis demonstrates that sets with low-quality sequences tend to increase this similarity.

Regarding Analysis 1.2, a higher number of initial reads was observed which initially suggested that the concatenation process may improve the final transcript. However, a superior number was expected since the concatenation of the assembled transcriptomes assures the representation of all samples' transcripts. Thus, it was not possible to determine with certainty whether the concatenation process improves or not the final reconstructed transcriptome. Similarly to the previous analysis, there was also a significant reduction in the number of final transcripts when comparing the pre- and post-evigene results. This analysis is referred to as *individual* in Table 4. Through the 'Num contigs > 200' statistic, it was possible to verify that approximately 84.5% of reads were eliminated from the initial set. This demonstrates that the EvidentialGene software was successful in removing highly similar and duplicated sequences from the initial set of contigs. The obtained reduction percentage was more elevated than any percentage of removed reads in Analysis 1.1. It is possible to infer a higher number of similar contigs when the assembled transcriptomes are concatenated. The predominance of highly similar sequences supports the low quality of the sequenced RNA.

Regarding the assessment of redundancy upon the transcriptome assembled during Analysis 2, a decrease in the final transcripts was also observed when comparing the pre- and post-evigene results. EvidentialGene revealed to be once again successful in removing highly similar and duplicated sequences from the initial set of reads. A superior number of initial reads was observed, once again suggesting that the concatenation process may improve the final transcript. However, a higher number was expected since merging all raw reads assures the representation of all sample sequences. Thus, it was also not possible to determine with certainty whether the concatenation process improves or not the final reconstructed transcriptome. Approximately 72.7% of contigs were eliminated, with the concatenation of all raw reads prior to the *de novo* assembly. This analysis is referred to as *all* in Table 4. This percentage was superior to any eliminated contigs in Analysis 1.1, though slightly inferior to the percentage of removed reads in Analysis 1.2. Once again, the high predominance of duplicated contigs supports the low quality of the sequenced RNA.

EVIGENE Analysis							
Sample	Max contig length	Num contigs >200	Total bases in contigs >200	N50 for contigs >200	Contigs >200 in N50	GC contigs >200	Mean length for contigs >200
<i>C10_pre_evigene</i>	8882	39349	30381505	1196	7422	41.7	772.1
<i>C10_pos_evigene</i>	6627	12242	6647595	645	2956	48.8	543.0
<i>C13_pre_evigene</i>	15539	66292	56321735	1473	10393	42.6	849.6
<i>C13_pos_evigene</i>	14502	20108	12871695	846	4223	49.3	640.1
<i>C3_pre_evigene</i>	5797	7659	3932912	609	1840	43.0	513.5
<i>C3_pos_evigene</i>	3192	2092	987582	501	605	50.1	472.1
<i>G1R1_pre_evigene</i>	16907	121976	103735844	1334	23928	42.6	850.5
<i>G1R1_pos_evigene</i>	14502	39304	22818426	705	9542	48.9	580.6
<i>R11_pre_evigene</i>	4710	11856	6471944	644	3351	40.0	545.9
<i>R11_pos_evigene</i>	3393	3042	1231404	420	943	47.6	404.8
<i>R12_pre_evigene</i>	15829	58448	51322188	1525	9223	42.5	878.1
<i>R12_pos_evigene</i>	14502	17793	11037357	798	3847	49.5	620.3
<i>R15_pre_evigene</i>	13696	38513	23054946	778	9247	41.3	598.6
<i>R15_pos_evigene</i>	13287	10146	4536462	477	2955	48.3	447.1
<i>R3_pre_evigene</i>	6977	13834	8199695	756	3231	40.6	592.7
<i>R3_pos_evigene</i>	6627	3731	1682391	480	1050	48.5	450.9
<i>individual_pre_evigene</i>	16907	357927	283420769	1227	64906	42.3	791.8
<i>individual_pos_evigene</i>	14502	55639	30281481	663	12446	48.3	544.2
<i>all_pre_evigene</i>	19307	193986	170953443	1464	33073	42.6	881.3
<i>all_pos_evigene</i>	14502	52989	32443191	792	11223	49.0	612.3

Table 4 – Evigene quality assessment metrics. Results obtained per sample, with (post-evigene) and without (pre-evigene) removal of redundant transcripts for comparison. *individual* and *all* nomenclatures concern the concatenated assemblies and merged raw data, respectively.

5. Assemblies completeness output

The final sets of reads (post-evigene) obtained for each conducted analysis were used to perform completeness assessments using BUSCO. A completeness evaluation was also conducted upon the sets of transcripts solely assembled for a pre- and post-evigene comparison. Tables 5 and 6 summarise the results in percentages obtained for the three conducted analyses using EvidentialGene, considering all four BUSCO classifications.

Regarding Analysis 1.1, the number of genes with only a single-copy slightly increased for all samples, when comparing the pre- and post-evigene results, with the exception of samples C10 and C13. The decrease observed for both these samples was small. The highest percentages of single-copy genes were exhibited by samples G1R1 and C13, prior and posterior to the removal of redundancy. The opposite occurred for samples C3 and R11. In relation to the number of genes appearing in duplicate, there was also a small rise for the majority of samples when comparing the pre- and post-evigene results, with the exception of samples R11 and R3. The reference sample R11 remained unaltered. The number of duplicated genes decreased 1.18% for the sample R3. In both pre- and post-evigene results, samples G1R1 and C13 presented a higher percentage of complete and duplicated genes. The opposite occurred for the sample C3, and the reference samples R3 and R11. Regarding the number of fragmented genes, a decrease was observed for most transcriptomes when comparing the pre- and post-evigene results. Samples C3 and R11 exhibited a non-significant increase of 0.39%, and sample R3 remained unaltered. Lastly, the percentage of genes not found in the transcriptomes decreased for all samples. The highest percentages of missing genes were exhibited by samples C3 and R11, prior and posterior to the removal of redundant sequences. The opposite occurred for the contaminated samples G1R1 and C13. The obtained results make it possible to still observe the impact of low-quality sequencing among all samples. The percentage of complete and single-copy genes was inferior to 50% for all samples for pre- and post-evigene analysis. Moreover, at least three reconstructed transcriptomes exhibited a very low percentage of these genes (less than 15%). The observed low-quality is also supported by a percentage of missing genes much superior to the percentage of single-copy genes for most samples. Additionally, the percentage of fragmented genes was also slightly superior for some transcriptomes. The identification of genes under single-copy control allows to infer the core genes of the last common ancestor of a given organism, as well as its associated protein sequences. This facilitates the

understanding of genes' primitive function, their role on a given metabolic pathway, and their responses to environmental stimuli and alterations. Contrarily, a higher number of missing genes in a transcriptome hampers gene classification and functional annotation, not allowing us to infer the main genes essential to an organism's survival. In the present work, these results suggest that samples G1R1 and C13 were better sequenced. The opposite occurred for samples C3 and R11.

Regarding Analysis 1.2, the number of single-copy genes increased when comparing the pre- and post-evigene results. A slightly increase can be observed for the percentage of fragmented and missing genes. The percentage of duplicated genes decreased. The concatenation of all assembled transcriptomes assures the representation of all samples transcripts and associated fragments and genes. In this case, compared to Analysis 1.1, this led to a superior and improved percentage of genes under single-copy control and a much more reduced number of missing genes. These results suggest that it would be more advantageous to use all assemblies concatenated, rather than each transcriptome reconstructed individually, as a reference for gene identification and annotation.

Regarding Analysis 2, a substantial increase of complete and single-copy genes was also observed, when comparing the pre- and post-evigene results. Similar to Analysis 1.2, the percentage of fragmented and missing genes slightly increased. The number of duplicated genes also decreased. Once more, the concatenation of all raw reads before the de novo assembly assures the representation of all samples' sequences and associated genes. In the present part, compared to Analysis 1.1, the concatenation of all RNA sequences led to a major improvement in the percentage of genes under single-copy control and a lower number of genes not found along the transcriptome. These results suggest that assembling a transcriptome from a set of concatenated reads leads to a more complete final transcript. Its use as a reference for gene annotation would be more advantageous and accurate than using each transcriptome assembled individually. Compared to Analysis 1.2, the results also seem to have improved, with an increase in the percentage of complete and single-copy genes and a decrease of missing genes. The obtained results suggest that assembling a transcriptome from a set of concatenated reads improves its completeness, facilitating the identification of all core genes in an organism.

The removal of redundant sequences led to an evident better-quality and improvement of results in Analyses 1.2 and 2. The percentage of complete and single-copy genes was superior to 60% for both analyses. Consequently, the percentage of

missing genes was inferior to 15%. Concatenate all assemblies prior to a redundancy removal, and reconstruct a transcriptome from merge raw reads, assures that a higher percentage of core genes is represented, improving completeness and gene identification and annotation. This way, it allows us to retrieve more reliable biological conclusions. To conclude, it is important to consider that BUSCO classifications are based on HMM score and length cut-offs. A gene with no BUSCO matches (missing) could be present as fragments that fail to align or do not meet the established thresholds. Complete and fragmented classifications can also be wrongly attributed to genes that do not meet BUSCO profile length and score expectations. Genes may be classified as fragmented when they are complete, or vice versa.

BUSCO Analysis (Pre-Evigene)				
	Complete Single-Copy	Complete Duplicate	Fragmented	Missing
<i>C10_pre_evigene</i>	34,51%	4,71%	23,53%	37,25%
<i>C13_pre_evigene</i>	45,10%	10,98%	24,71%	19,22%
<i>C3_pre_evigene</i>	1,96%	0,00%	5,88%	92,16%
<i>G1R1_pre_evigene</i>	41,18%	23,14%	18,04%	17,65%
<i>R11_pre_evigene</i>	3,53%	0,00%	10,98%	85,49%
<i>R12_pre_evigene</i>	36,47%	9,02%	27,45%	27,06%
<i>R15_pre_evigene</i>	14,90%	2,35%	22,35%	60,39%
<i>R3_pre_evigene</i>	4,71%	1,18%	11,37%	82,75%
<i>individual_pre_evigene</i>	14,90%	63,53%	14,12%	7,45%
<i>all_pre_evigene</i>	43,53%	47,06%	7,06%	2,35%

Table 5 – BUSCO assessment metrics. Results obtained per sample, with no removal (pre-evigene) of redundant transcripts. *individual* and *all* nomenclatures concern the concatenated assemblies and merged raw data, respectively.

BUSCO Analysis (Post-Evigene)				
	Complete Single-Copy	Complete Duplicate	Fragmented	Missing
<i>C10_pos_evigene</i>	32,94%	8,63%	23,14%	35,29%
<i>C13_pos_evigene</i>	44,31%	14,90%	22,75%	18,04%
<i>C3_pos_evigene</i>	3,14%	0,39%	6,27%	90,20%
<i>G1R1_pos_evigene</i>	42,75%	26,67%	15,29%	15,29%
<i>R11_pos_evigene</i>	3,92%	0,00%	11,37%	84,71%
<i>R12_pos_evigene</i>	42,75%	11,76%	22,75%	22,75%
<i>R15_pos_evigene</i>	17,25%	5,49%	18,82%	58,43%
<i>R3_pos_evigene</i>	6,67%	0,00%	11,37%	81,96%
<i>individual_pos_evigene</i>	61,18%	10,20%	16,47%	12,16%
<i>all_pos_evigene</i>	74,12%	11,76%	8,24%	5,88%

Table 6 - BUSCO assessment metrics. Results obtained per sample, with removal (post-evigene) of redundant transcripts. *individual* and *all* nomenclatures concern the concatenated assemblies and merged raw data, respectively.

6. Overall discussion of the results

The discrepancy and low-quality among all collected samples were evident along the present work. This may have occurred due to errors in the sequencing process. Since all established methodological analyses were conducted upon low-quality sequences of RNA, it was not possible to retrieve reliable biological conclusions. The preliminary analysis of quality control and trimming, performed upon the sequenced raw reads, allowed to firstly observe a disparity in the number of initial sequences for all samples. A sequencing process with few errors would provide a similar number of raw reads. Though the trimming step ensured elimination of reads with low-score nucleotides and inserted adapters, this was insufficient to improve the quality of the samples' sequences. The results obtained with the assessment of quality using FastQC also suggested a biased library and problems in the sequencing process. Moreover, they also indicated a high number of reads presenting duplicates for each sample.

The *de novo* assembly approach was conducted upon the trimmed sequences of each collected sample. The statistics provided to evaluate the overall quality of the reconstructed assemblies supported the predominance of discrepancy. The assessment of reads representation and the number of assembled transcripts demonstrated an evident non-uniform distribution among samples. In both assessments, samples G1R1 and C13 exhibit the highest values. These may suggest that they were better sequenced. However, alternative hypotheses can be considered, for instance, contamination from microorganisms, such as parasites, commonly found in the *Carcinus maenas* digestive gland. In parallel, the same assembly strategy was conducted upon all raw sequences of each sample concatenated into a single file. The objective of performing a second analysis was to verify if merging all raw reads prior to a *de novo* assembly would lead to a better reconstructed transcript. In other words, improve the number of reconstructed contigs and the length of the final fragment. In the present work, the process of concatenation appears to have improved the final assembled transcript. To determine the levels of disparity and low-quality among the assembled sequences of each sample, three computational analyses were conducted: a PCA, a MA plot, and a Dendrogram and Heatmap. The *Principal Component Analysis* evidenced a complete isolation of the sample G1R1, supporting its previous observed contribution for a high discrepancy. The remaining samples clustered in three groups, given its similarity in counts of total assembled reads and total assembled transcripts. The MA plot suggested similar expression patterns for the two treatment groups (Minho and Lima estuaries). Moreover, it highlighted genes with highly

differential expression levels. Lastly, the dendrogram and heatmap supported the presence of genes that are more and less expressed in organisms collected from the Lima estuary compared with the Minho estuary. Despite the particular suggestions of each plot, an absence of coherence among all samples was still visible and supported by each computational analysis.

Highly similar transcripts may occur when assembling a transcriptome. Its occurrence may be increased among sequences with lower quality. For instance, insufficient and incomplete reads. Since transcripts are commonly over-sequenced, large sets of duplicates are frequently generated in the case of transcriptomes. For this reason, a redundancy assessment was conducted upon the assembled transcriptomes, using EvidentialGene. In order to observe the effects of minimising redundancy among the sets of contigs, it was established a comparison between transcripts submitted to the process of redundancy removal and transcripts solely assembled. When assessing each individual transcriptome, a significant reduction in the number of final transcripts was observed, when comparing the pre- and post-evigene results. This suggested that the EvidentialGene software, and its CDS approach, successfully removed similar and duplicated sequences from the sets of contigs. The elimination of a high number of duplicated sequences suggested once more the low-quality of the RNA sequences. However, alternative hypotheses can be considered, for instance, the occurrence of alternative splicing. Moreover, this analysis demonstrated that sets with low-quality contigs tend to increase similarity between sequences. Besides conducting a redundancy assessment upon each sample's transcriptome, another evaluation and removal of redundant sequences was also performed upon all assemblies concatenated into a single file, to determine whether the process of concatenation would result in a better reconstructed transcript. The CDS approach, implemented by Evigene, also successfully reduced the number of duplicated sequences. Nevertheless, highly similar sequences were still very present among the set, suggesting the low quality of the sequenced RNA. A higher number of similar contigs was observed with the concatenation of the assembled transcriptomes, which primarily suggested that the concatenation process may improve the final transcript. However, a superior number was expected since the concatenation of the assembled transcriptomes assures the representation of all samples' transcripts. Thus, it was not possible to determine with certainty whether the concatenation process improves or not the final reconstructed transcriptome. The assessment conducted upon the transcript reconstructed from the prior concatenated raw reads also demonstrated a decrease in the final transcripts, when comparing the pre- and post-evigene results. EvidentialGene revealed to be once

again successful in removing highly similar and duplicated sequences from the initial set of reads. A superior number of initial reads was observed, since merging all raw reads assures the representation of all sample sequences. Thus, it was also not possible to determine with certainty whether the concatenation process improves or not the final reconstructed transcriptome. Once again, the high predominance of duplicated contigs suggested the low quality of the sequenced RNA, or the occurrence of biological processes such as alternative splicing.

To conclude the methodology work, an assessment of completeness was performed using the BUSCO tool. In this analysis, the BUSCO program was conducted upon transcripts submitted to the process of redundancy removal, and transcripts solely assembled to establish a comparison. The number of single-copy genes slightly increased for the majority of samples. The percentage of missing genes decreased for all samples. Samples G1R1 and C13 exhibited the highest percentages of single-copy genes, suggesting that they were better sequenced. The opposite occurred for samples C3 and R11. However, it was still possible to observe a predominance of disparity and low-quality among all samples. This was supported by a percentage of complete and single-copy genes inferior to 50% for all samples, transcriptomes exhibiting less than 15% of single-copy genes, and a percentage of missing genes much superior to the percentage of single-copy genes for most samples. BUSCO proved to be successful in assessing the completeness of transcriptomes. In the case of the concatenated assemblies, there was an improvement in the completeness results. The number of single-copy genes suffered a major increase. Since the process of concatenation assures the representation of all samples, and associated fragments and genes, a superior and improved percentage of genes under single-copy control, and a much more reduced number of missing genes, were observed. This suggests that it would be more advantageous to use all assemblies concatenated, rather than each transcriptome reconstructed individually, as a reference for gene identification and annotation. In this analysis, BUSCO also proved to be successful in assessing the completeness of transcriptomes. In relation to the transcript reconstructed from the prior concatenated raw reads, there was also an improvement in the completeness results. A substantial increase of complete and single-copy genes was equally observed, as well as a much inferior number of missing genes along the transcriptome. When compared to both previous analyses, the results suggested that assemble a transcriptome from a set of concatenated reads leads to a more complete final transcript. Its use as a reference for gene annotation would be more advantageous and accurate, facilitating the identification of all core genes in an organism.

Chapter IV. Conclusions

1. Final Remarks

The increase of pollutants discharge along estuarine ecosystems requires improved assessment and monitoring approaches to conserve estuaries' inhabiting organisms, local biodiversity and functioning. The daily exposure to stressors and contaminants leads to irremediable damage at organisms' lower organizational levels, inducing differential sensitivity, as well as altered biochemical and physiological responses. The use of biomarkers, particularly transcriptomic responses, has been useful to characterise the resultant effects from exposure to single contaminants or their cocktails. Their use has been particularly successful in the often-neglected crustaceans. Since the Crustacea sub-phylum shares significant similarity with the insect taxon, these is preferred by the scientific community (Chessa 2021; Nong et al. 2020). This negligence limits transcriptomic information on the crustaceans taxonomic group. In the Crustacea sub-phylum, the green crab *Carcinus maenas* has proven to be an adequate experimental model to investigate the effects of environmental contaminants at a transcriptomic level, and assess the health status of estuarine ecosystems. Biochemical alterations in *C. maenas*, when exposed to even moderate pollution, demonstrate sensitivity to abiotic and chemical stress, which is an advantage to understand the impact of environmental contamination. Future ecotoxicological research would beneficiate with a higher inclusion of these macroinvertebrates. The present thesis enhances *C.maenas* as a suitable species to be maintained in the laboratory, and to extract RNA for sequencing and further bioinformatic analyses. Crabs individuals were collected from the north-western Portuguese Lima and Minho estuaries, since they are under different anthropogenic pressure. The Lima river is under moderate contamination and the Minho river is under lower levels of human intervention. The selection of both these estuaries was also due to the differential sensitivity to chemical contamination and abiotic stress observed in their green crab populations, suggesting that the exposure history in Minho and Lima estuaries may influence their response to contamination (Rodrigues, Santos, et al. 2014; Rodrigues, Gravato, and Guimaraes 2013; Rodrigues et al. 2015; Rodrigues et al. 2012). Furthermore, previous results suggest such differential responses occur via physiological acclimation, rather than adaptation.

The present study also highlights the use of bioinformatic and computational approaches as adequate to analyse high-throughput transcriptomic data. Despite the errors in the sequencing process, carried out in previous research, following Illumina TrueSeq protocol, the computational softwares applied along this thesis allowed to analyze sequences of RNA. Regarding the preliminary analysis of the initial generated raw sequence data, it was possible to conclude that the FastQC and Trimmomatic softwares were successful in reporting the quality of each extracted sample, and removing low score/poor quality reads and adapters inserted for the molecular sequencing, respectively. This lead to the conclusion that pre-processing NGS data (sequences quality control; read filtering and trimming) minimizes erroneous inferences in biological interpretations. The present thesis also supported that the Trimmomatic tool detects adapters at both reads extremities (paired-end approach).

In relation to the assembly step, it was possible to conclude that the *de novo* approach is adequate to generate full-length transcripts, when no reference transcriptome is available. The Trinity assembler was successful in reconstructing transcriptomes solely through short RNA-Seq reads. However, the *de novo* strategy demonstrated to be susceptible to errors introduced during the sequencing process. Moreover, this approach hardly differentiated highly similar sequences, assembling them into single transcripts. Concerning the second analysis conducted upon all raw sequences concatenated, it was possible to verify an improvement in the number of reconstructed contigs and the length of the final fragment. In other words, the process of concatenation appeared to improve the final assembled sequence. However, it is not possible to conclude that this process will always result in a better full-length transcript, since previous research has demonstrated the opposite (Chessa 2021).

When assembling a transcriptome particularly from reads with lower quality, the occurrence of highly similar transcripts tends to increase. Moreover, since transcripts are commonly over-sequenced, it is common to observe redundancy and duplicates among reads. The present thesis supported that sets with low-quality sequences tend to increase this similarity. It was also possible to conclude that the EvidentialGene software, and its CDS approach, is successful in eliminating similar and duplicated sequences from a set. Concerning the analyses conducted upon all assemblies concatenated and the transcript reconstructed from the prior merged raw reads, the present study revealed a superior number of initial reads, since the process of concatenation assures the representation of all samples' transcripts. Thus, it was not possible to determine with certainty whether the concatenation process improves or not the final reconstructed transcriptome.

Regarding the completeness quantification of the assembled transcriptomes, it was possible to conclude that the BUSCO tool successfully searches for genes only present as single-copy orthologs in a given lineage or taxonomic group, including the eukaryotes. It was also possible to verify that BUSCO facilitates these genes identification since it separates genes into four main classifications: 1. Complete, subdivided into 'Complete and single-copy' and 'Complete and duplicated', 2. Fragmented, 3. Missing. Concerning the analyses conducted upon all assemblies concatenated and the transcript reconstructed from the prior merged raw reads, the present study revealed that assemble a transcriptome from a set of concatenated reads leads to a more complete final transcript. Its use as a reference for gene annotation would be more advantageous and accurate, facilitating the identification of all core genes in the green crab. Future research in gene identification and annotation would benefit with the process of concatenation prior to the assembly performance.

2. Future Work

The work developed in this study is a starting point to characterize the core genes identified in the green crab *Carcinus maenas*, as well as the metabolic pathways in which they are involved. Since crabs individuals from estuarine environments with distinct levels of contamination are being considered, future research should focus on researching the core genes with differential responses, leading to biochemical and physiological alterations. The full-length transcript reconstructed from the set of concatenated reads should be used as the model assembly (reference). This way, further steps should involve:

1. performing a Blast against a non-redundant metazoan database to assess the primary function of these genes;
2. obtaining Gene Ontology (GO) classifications. Assign GO terms to describe gene activities occurring at a molecular, cellular or biological level;
3. using the KEGG pathway database to understand genes interactions and relations;
4. identifying up- and down-regulated genes;
5. establishing a comparison between the most relevant up- and down-regulated genes in crabs from the estuary with moderate contamination and crabs from the low-impacted estuary;
6. employing the DEBseq package to statistically analyse possible differences in gene transcripts between crabs of the two estuaries.

Since a sequenced, assembled and annotated green crab transcriptome became available in the (Verbruggen et al. 2015) article, further research may be conducted using this genome as a reference.

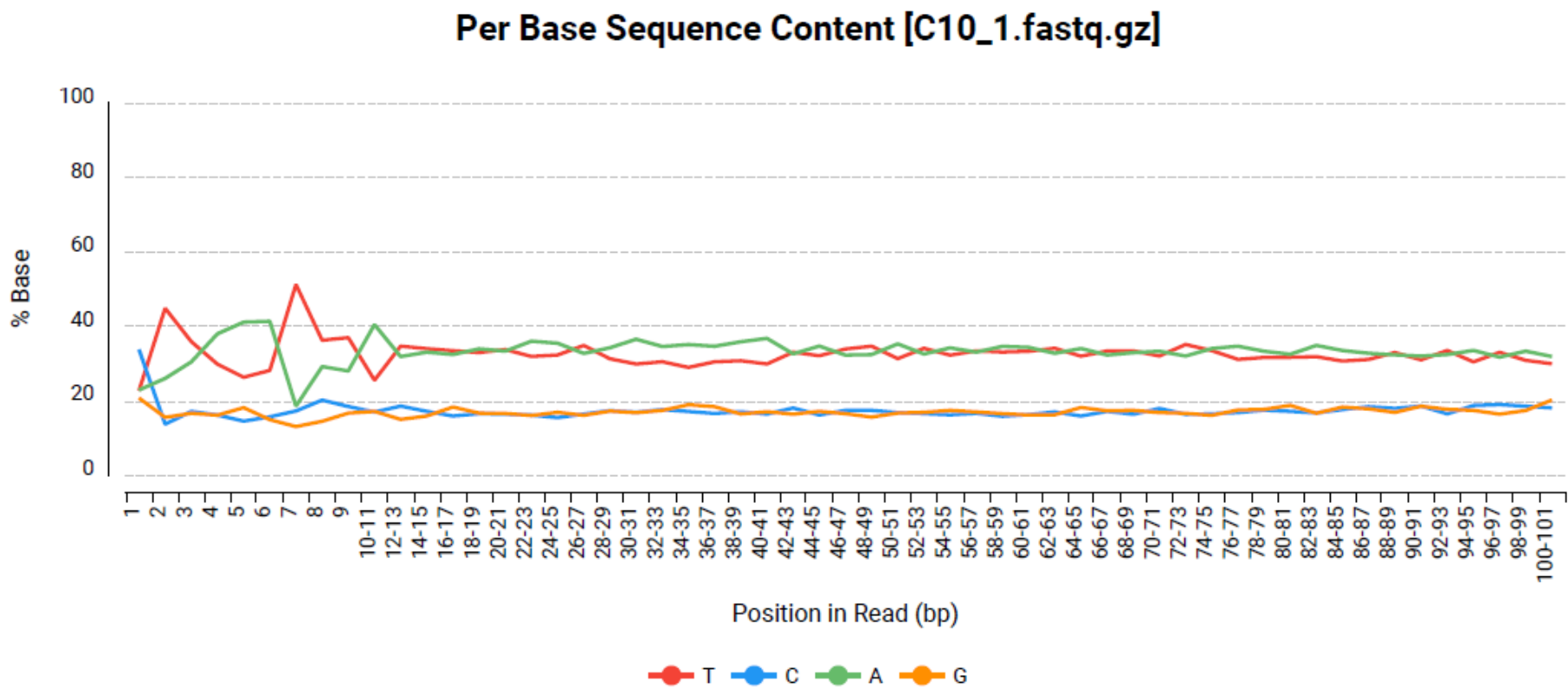
The availability of new-sequenced and assembled transcriptomes for the Crustacea sub-phylum may increase the genomic and transcriptomic information of its species, not limiting their use in scientific research. The area of ecotoxicology would benefit with the increase of information on crustaceans, particularly the crab *Carcinus maenas*.

Appendix

The Appendix section contains all the Supplementary Figures generated to support the obtained Results. It comprises the following: representation of FastQC module plots for two contaminated samples.

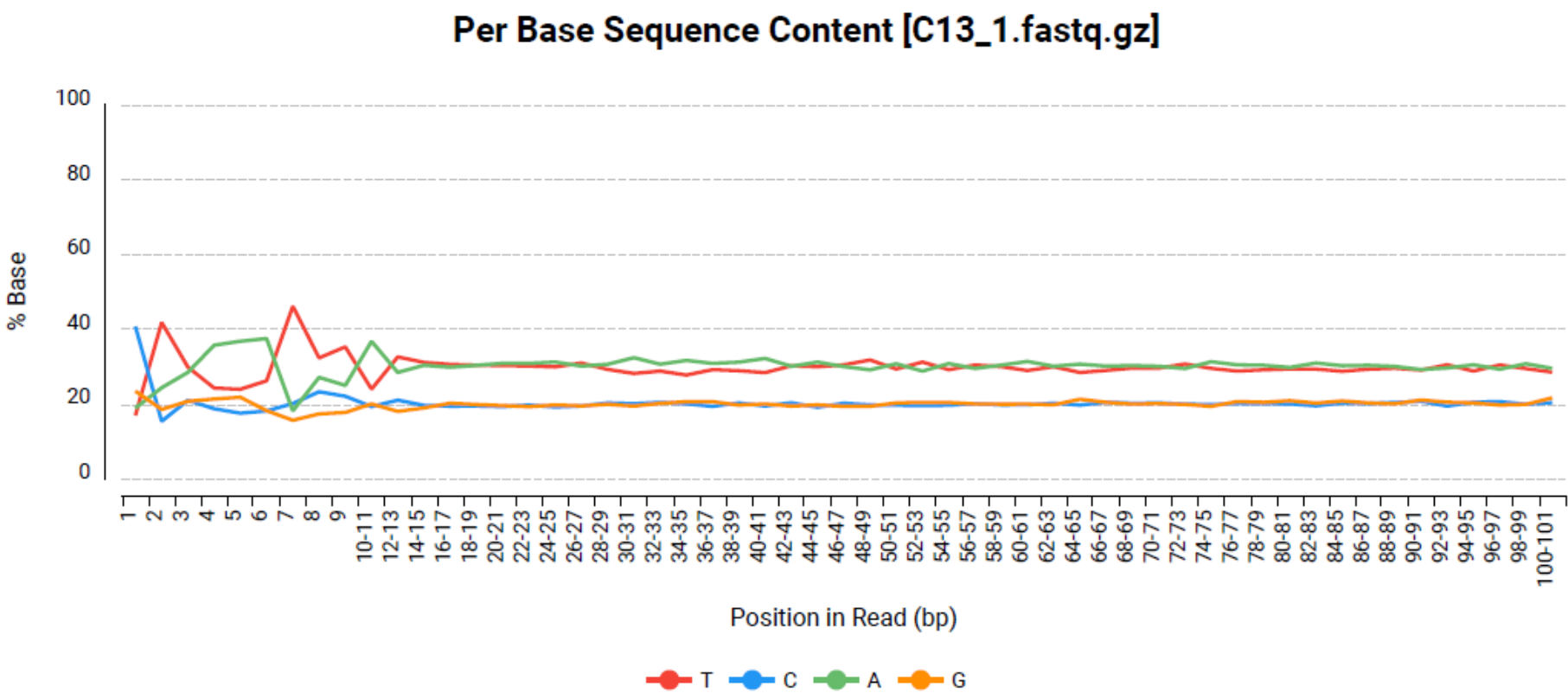
Supplementary Figure 1

Supplementary Figure 1 – ‘Per Base Sequence Content’ module plot, for sample C10. Percentage of base pairs, and its read positions are identified.



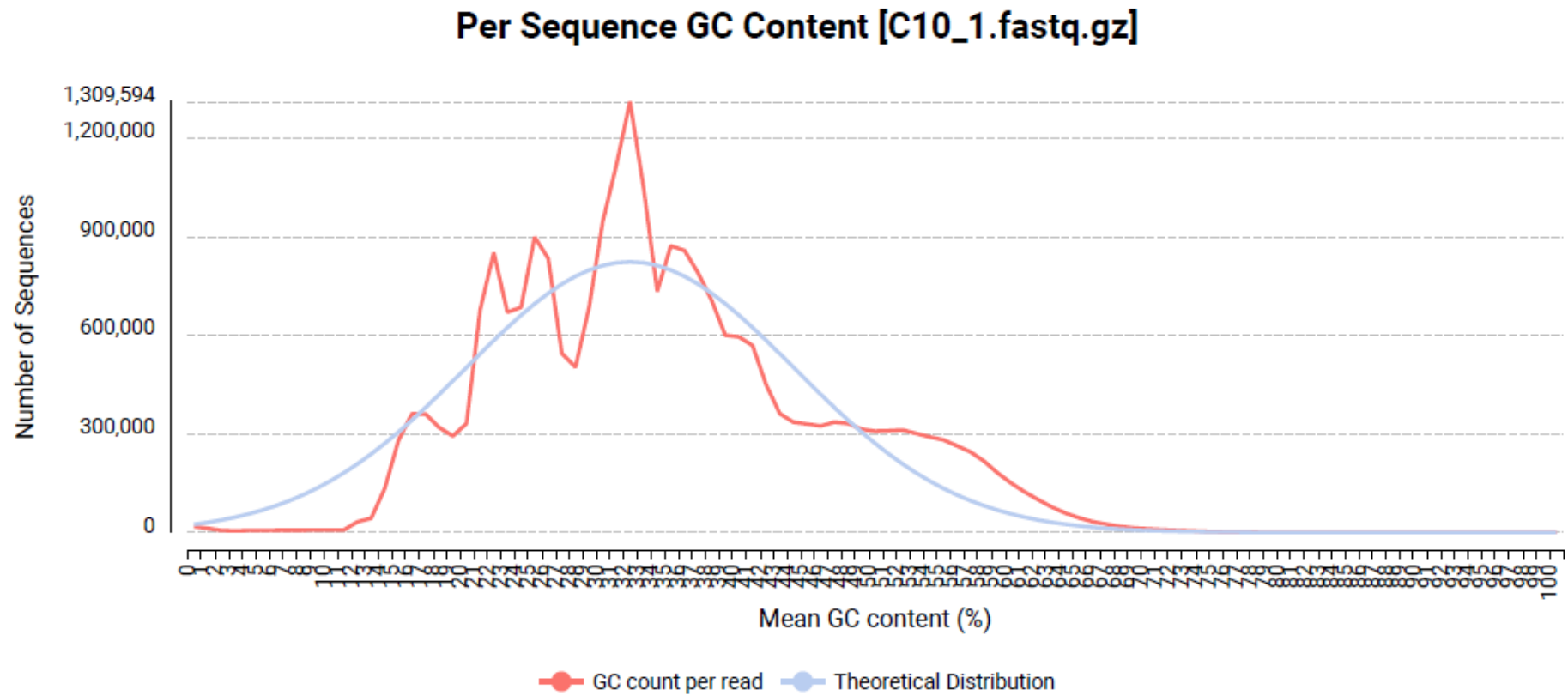
Supplementary Figure 2

Supplementary Figure 2 – ‘Per Base Sequence Content’ module plot, for sample C13. Percentage of base pairs, and its read positions are identified.



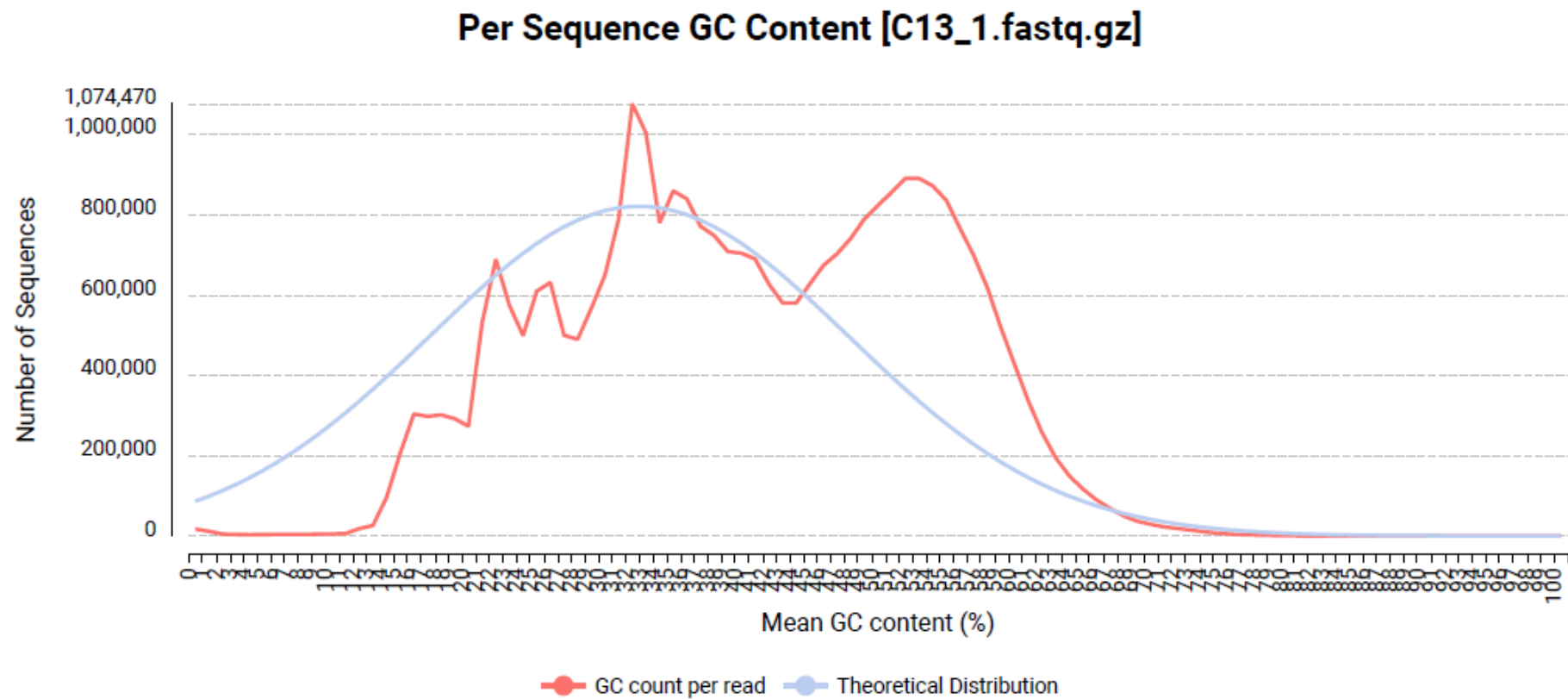
Supplementary Figure 3

Supplementary Figure 3 – 'Per Sequence GC Content' module plot, for sample C10. Percentage of mean GC content per number of sequences is identified (red line). A theoretical distribution (blue line) is represented for comparison.



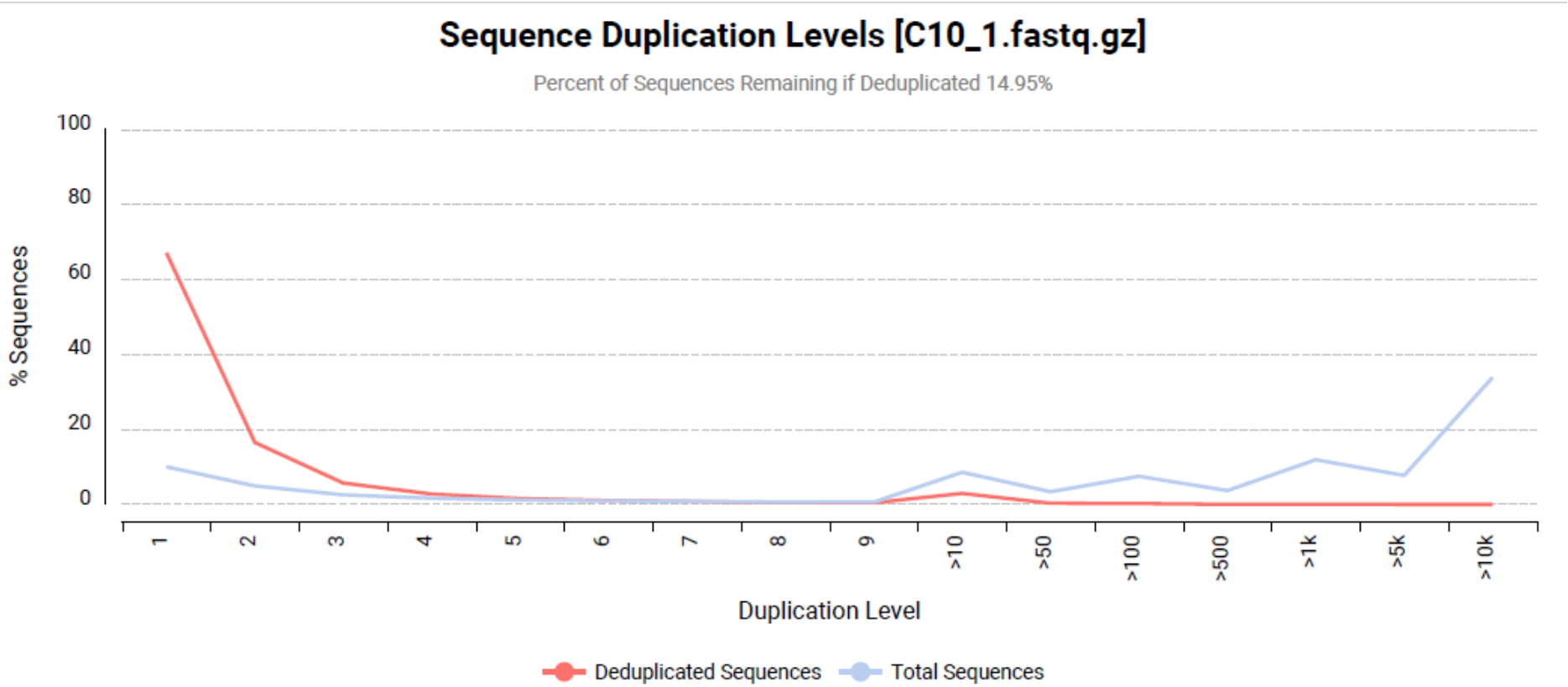
Supplementary Figure 4

Supplementary Figure 4 – 'Per Sequence GC Content' module plot, for sample C13. Percentage of mean GC content per number of sequences is identified (red line). A theoretical distribution (blue line) is represented for comparison.



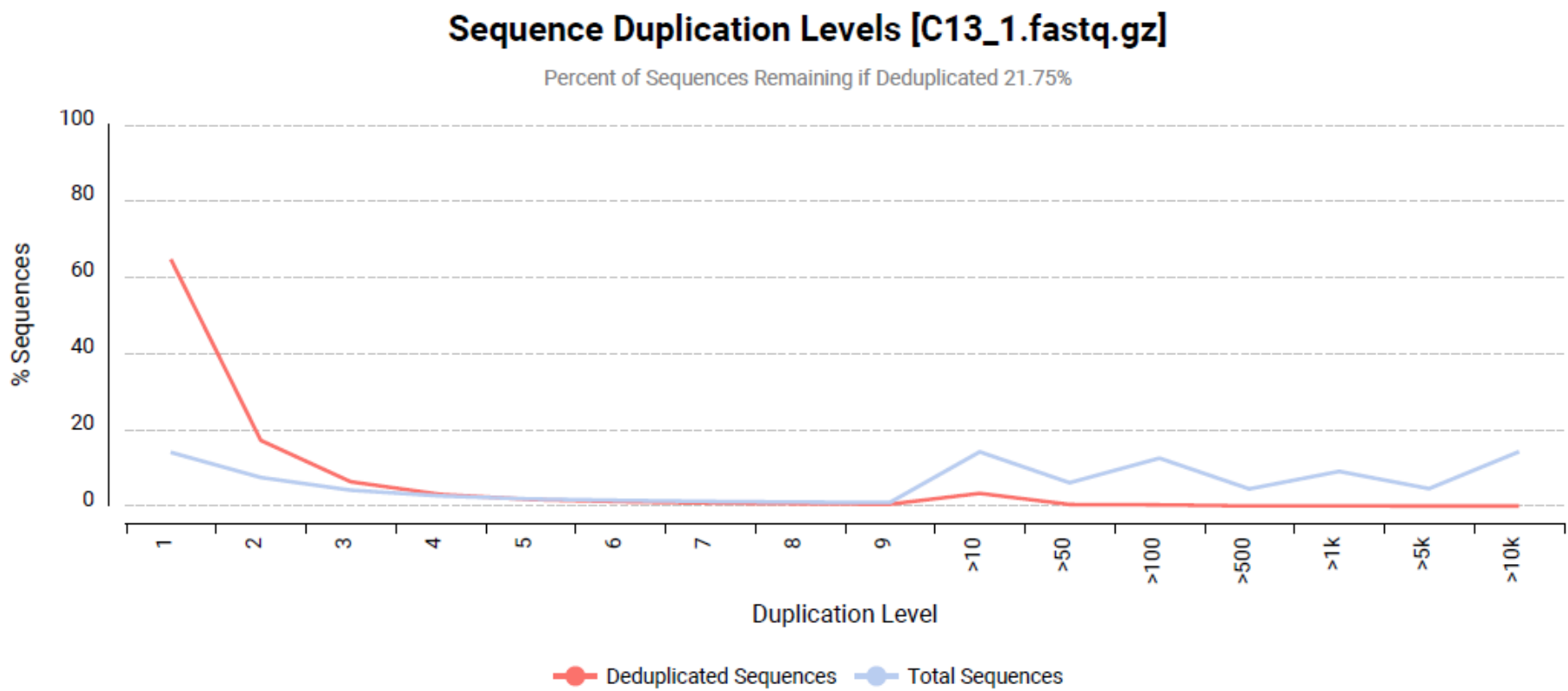
Supplementary Figure 5

Supplementary Figure 5 – ‘Sequence Duplication Levels’ plot, for sample C10. Level of duplication per percentage of sequences is identified (red line). Total sequences (blue line) is represented for comparison.



Supplementary Figure 6

Supplementary Figure 6 – ‘Sequence Duplication Levels’ plot, for sample C13. Level of duplication per percentage of sequences is identified (red line). Total sequences (blue line) is represented for comparison.



Bibliography

- Abdi, Hervé, and Lynne J Williams. 2010. 'Principal component analysis', Wiley interdisciplinary reviews: computational statistics, 2: 433-459.
- Alhakami, Hind, Hamid Mirebrahim, and Stefano Lonardi. 2017. 'A comparative evaluation of genome assembly reconciliation tools', Genome Biology, 18: 93.
- Alkan, Can, Saba Sajjadian, and Evan E Eichler. 2011. 'Limitations of next-generation genome sequence assembly', Nature methods, 8: 61-65.
- Amiard, J-C, C Amiard-Triquet, Sabria Barka, Jocelyne Pellerin, and PS Rainbow. 2006. 'Metallothioneins in aquatic invertebrates: their role in metal detoxification and their use as biomarkers', Aquatic toxicology, 76: 160-202.
- Ansorge, Wilhelm J. 2009. 'Next-generation DNA sequencing techniques', New biotechnology, 25: 195-203.
- Antunes, C, and S Costa Dias. 2005. 'The faunal composition of the euryhaline section of the Lima River estuary', Ichthyology c/c of voprosy ikhtiologii, 45: 163.
- Azevedo, Isabel, Sandra Ramos, Ana Paula Mucha, and Adriano A Bordalo. 2013. 'Applicability of ecological assessment tools for management decision-making: a case study from the Lima estuary (NW Portugal)', Ocean & coastal management, 72: 54-63.
- Bamber, SD, and MH Depledge. 1997. 'Responses of shore crabs to physiological challenges following exposure to selected environmental contaminants', Aquatic Toxicology, 40: 79-92.
- Bardou, Philippe, Jérôme Mariette, Frédéric Escudié, Christophe Djemiel, and Christophe Klopp. 2014. 'jvenn: an interactive Venn diagram viewer', BMC bioinformatics, 15: 1-7.
- Bentley, David R, Shankar Balasubramanian, Harold P Swerdlow, Geoffrey P Smith, John Milton, Clive G Brown, Kevin P Hall, Dirk J Evers, Colin L Barnes, and Helen R Bignell. 2008. 'Accurate whole human genome sequencing using reversible terminator chemistry', Nature, 456: 53-59.
- BioBam, Bioinformatics. 2019. 'OmicsBox - Bioinformatics Made Easy'. <https://www.biobam.com/omicsbox/>.
- Birol, Inanç, Shaun D Jackman, Cydney B Nielsen, Jenny Q Qian, Richard Varhol, Greg Stazyk, Ryan D Morin, Yongjun Zhao, Martin Hirst, and Jacqueline E Schein. 2009. 'De novo transcriptome assembly with ABySS', Bioinformatics, 25: 2872-2877.
- Bolger, Anthony M, Marc Lohse, and Bjoern Usadel. 2014. 'Trimmomatic: a flexible trimmer for Illumina sequence data', Bioinformatics, 30: 2114-2120.
- Bowe, Alexander, Taku Onodera, Kunihiko Sadakane, and Tetsuo Shibuya. 2012. 'Succinct de Bruijn graphs', International workshop on algorithms in bioinformatics, 225-235.
- Brown, C Titus, Adina Howe, Qingpeng Zhang, Alexis B Pyrkosz, and Timothy H Brom. 2012. 'A reference-free algorithm for computational normalization of shotgun sequencing data', ArXiv, 1203-4802.
- Brown, Joseph, Meg Pirrung, and Lee Ann McCue. 2017. 'FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool', Bioinformatics, 33: 3137-3139.
- Buckingham, Lela. 2019. Molecular diagnostics: fundamentals, methods and clinical applications, FA Davis.
- Cairrao, E, M Couderchet, AMVM Soares, and L Guilhermino. 2004. 'Glutathione-S-transferase activity of Fucus spp. as a biomarker of environmental contamination', Aquatic Toxicology, 70: 277-286.

- Carew, Melissa E, Vincent J Pettigrove, Leon Metzeling, and Ary A Hoffmann. 2013. 'Environmental monitoring using next generation sequencing: rapid identification of macroinvertebrate bioindicator species', *Frontiers in zoology*, 10: 1-15.
- Carlisle, DB, and LM Passano. 1953. 'The X-organ of Crustacea', *Nature*, 171: 1070-1071.
- Casneuf, Tineke, Yves Van de Peer, and Wolfgang Huber. 2007. 'In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation', *BMC bioinformatics*, 8: 1-13.
- Chaisson, Mark J, Dumitru Brinza, and Pavel A Pevzner. 2009. 'De novo fragment assembly with short mate-paired reads: Does the read length matter?', *Genome research*, 19: 336-346.
- Chang, Ernest S, Rainer Keller, and Sharon A Chang. 1998. 'Quantification of crustacean hyperglycemic hormone by ELISA in hemolymph of the lobster, *Homarus americanus*, following various stresses', *General and comparative endocrinology*, 111: 359-366.
- Chang, Ernest S, and Donald L Mykles. 2011. 'Regulation of crustacean molting: a review and our perspectives', *General and comparative endocrinology*, 172: 323-330.
- Chang, Ernest S, Glenn D Prestwich, and Marilyn J Bruce. 1990. 'Amino acid sequence of a peptide with both molt-inhibiting and hyperglycemic activities in the lobster, *Homarus americanus*', *Biochemical and Biophysical Research Communications*, 171: 818-826.
- Chariton, Anthony A, Leon N Court, Diana M Hartley, Matthew J Colloff, and Christopher M Hardy. 2010. 'Ecological assessment of estuarine sediments by pyrosequencing eukaryotic ribosomal DNA', *Frontiers in Ecology and the Environment*, 8: 233-238.
- Chen, Chuming, Sari S Khaleel, Hongzhan Huang, and Cathy H Wu. 2014. 'Software for pre-processing Illumina next-generation sequencing short read sequences', *Source code for biology and medicine*, 9: 1-11.
- Chen, Hsiang-Yin, Jean-Yves Toullec, and Chi-Ying Lee. 2020. 'The crustacean hyperglycemic hormone superfamily: progress made in the past decade', *Frontiers in Endocrinology*, 11: 769.
- Chen, S., J. S. McElroy, F. Dane, and E. Peatman. 2015. 'Optimizing Transcriptome Assemblies for Eleusine indica Leaf and Seedling by Combining Multiple Assemblies from Three De Novo Assemblers', *Plant Genome*, 8: 1-10.
- Chen, Shifu, Tanxiao Huang, Yanqing Zhou, Yue Han, Mingyan Xu, and Jia Gu. 2017. 'AfterQC: automatic filtering, trimming, error removing and quality control for fastq data', *BMC bioinformatics*, 18: 91-100.
- Chessa, Adriano Romário Correia. 2021. 'Effects of TBT on gene expression, locomotory behaviour and protein affinity prediction of *Triops longicaudatus*'. [Master's thesis, Instituto de Ciências Biomédicas Abel Salazar do Porto] <https://hdl.handle.net/10216/137376>.
- Chevreaux, Bastien, Thomas Pfisterer, Bernd Drescher, Albert J Driesel, Werner EG Müller, Thomas Wetter, and Sándor Suhai. 2004. 'Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs', *Genome research*, 14: 1147-1159.
- Chia, Fu-Shiang, John Buckland-Nicks, and Craig M Young. 1984. 'Locomotion of marine invertebrate larvae: a review', *Canadian Journal of Zoology*, 62: 1205-1222.
- Chikhi, Rayan, Antoine Limasset, Shaun Jackman, Jared T Simpson, and Paul Medvedev. 2014. 'On the representation of de Bruijn graphs', *International conference on Research in computational molecular biology*, 35-55.
- Chikhi, Rayan, and Paul Medvedev. 2013. 'Informed and automated k-mer size selection for genome assembly', *Bioinformatics*, 30: 31-37.

- Chung, J Sook, Heinrich Dircksen, and Simon G Webster. 1999. 'A remarkable, precisely timed release of hyperglycemic hormone from endocrine cells in the gut is associated with ecdysis in the crab *Carcinus maenas*', *Proceedings of the National Academy of Sciences*, 96: 13103-13107.
- Chung, J Sook, N Zmora, H Katayama, and N Tsutsui. 2010. 'Crustacean hyperglycemic hormone (CHH) neuropeptides family: functions, titer, and binding to target tissues', *General and comparative endocrinology*, 166: 447-454.
- Clarke, Kaitlin, Yi Yang, Ronald Marsh, and LingLin Xie. 2013. 'Comparative analysis of de novo transcriptome assembly', *Science China Life Sciences*, 56: 156-162.
- Cocquet, Julie, Allen Chong, Guanglan Zhang, and Reiner A Veitia. 2006. 'Reverse transcriptase template switching and false alternative transcripts', *Genomics*, 88: 127-131.
- Compeau, Phillip EC, Pavel A Pevzner, and Glenn Tesler. 2011. 'How to apply de Bruijn graphs to genome assembly', *Nature biotechnology*, 29: 987-991.
- Consortium, International Human Genome Sequencing. 2004. 'Finishing the euchromatic sequence of the human genome', *Nature*, 431: 931-945.
- Costa-Dias, Sérgia, Ronaldo Sousa, and Carlos Antunes. 2010. 'Ecological quality assessment of the lower Lima Estuary', *Marine Pollution Bulletin*, 61: 234-239.
- Crothers, JH. 1967. 'The biology of the shore crab *Carcinus maenas* (L.) 1. The background-anatomy, growth and life history', *Field studies*, 2: 407-434.
- Danecek, Petr, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, and Robert M Davies. 2021. 'Twelve years of SAMtools and BCFtools', *Gigascience*, 10.
- DeFur, Peter L. 2004. 'Use and role of invertebrate models in endocrine disruptor research and testing', *Ilar*, 45: 484-493.
- Del Fabbro, Cristian, Simone Scalabrin, Michele Morgante, and Federico M Giorgi. 2013. 'An extensive evaluation of read trimming effects on Illumina NGS data analysis', *PLoS One*, 8.
- Denoeud, France, Jean-Marc Aury, Corinne Da Silva, Benjamin Noel, Odile Rogier, Massimo Delledonne, Michele Morgante, Giorgio Valle, Patrick Wincker, and Claude Scarpelli. 2008. 'Annotating genomes with massive-scale RNA sequencing', *Genome Biology*, 9: 1-12.
- Dias, Ester, Pedro Morais, Mardik Leopold, Joana Campos, and Carlos Antunes. 2012. 'Natural born indicators: great cormorant *Phalacrocorax carbo* (Aves: Phalacrocoracidae) as monitors of river discharge influence on estuarine ichthyofauna', *Sea Research*, 73: 101-108.
- Dohm, Juliane C, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. 2008. 'Substantial biases in ultra-short read data sets from high-throughput DNA sequencing', *Nucleic acids research*, 36.
- dos Santos, Antonina, A Miguel P Santos, David VP Conway, Cátia Bartilotti, Patrícia Lourenço, and Henrique Queiroga. 2008. 'Diel vertical migration of decapod larvae in the Portuguese coastal upwelling ecosystem: implications for offshore transport', *Marine Ecology Progress Series*, 359: 171-183.
- Doyle, K. 1996. 'The Source of Discovery: Protocols and Applications Guide, Promega', Madison, Wis, USA.
- Duan, Jialei, Chuan Xia, Guangyao Zhao, Jizeng Jia, and Xiuying Kong. 2012. 'Optimizing de novo common wheat transcriptome assembly using short-read RNA-Seq data', *BMC genomics*, 13: 1-12.

- Dufayard, Jean-François, Laurent Duret, Simon Penel, Manolo Gouy, François Rechenmann, and Guy Perrière. 2005. 'Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases', *Bioinformatics*, 21: 2596-2603.
- Dziegiel, Piotr. 2004. 'Expression of metallothioneins in tumor cells', *Pol J pathol*, 55: 3-12.
- Eddy, Sean R. 2011. 'Accelerated profile HMM searches', *PLoS computational biology*, 7.
- Espie, Patrick J, and John C Roff. 1995. 'Characterization of chitobiase from *Daphnia magna* and its relation to chitin flux', *Physiological Zoology*, 68: 727-748.
- Andrews, Simon. 2010. 'Babraham bioinformatics - FastQC a quality control tool for high throughput sequence data', <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
- Ferreira, JG. 2003. 'Identification of sensitive areas and vulnerable zones in transitional and coastal portuguese systems: application of the United States National Estuarine Eutrophication Assessment to the Minho, Lima, Douro, Ria de Aveiro, Mondego, Tagus, Sado, Mira, Ria Formosa and Guadiana systems'.
- Ferreira, Nuno GC, Rui G Morgado, Luís Cunha, Marta Novo, Amadeu MVM Soares, Andrew J Morgan, Susana Loureiro, and Peter Kille. 2019. 'Unravelling the molecular mechanisms of nickel in woodlice', *Environmental research*, 176: 108507.
- Fingerman, M, NC Jackson, and R Nagabhushanam. 1998. 'Hormonally-regulated functions in crustaceans as biomarkers of environmental pollution', *Comparative Biochemistry and Physiology Part C: Pharmacology, Toxicology and Endocrinology*, 120: 343-350.
- Fingerman, Milton, Wayne E Julian, Morris A Spirtes, and Richard M Kostrzewa. 1974. 'The presence of 5-hydroxytryptamine in the eyestalks and brain of the fiddler crab *Uca pugilator*, its quantitative modification by pharmacological agents, and possible role as a neurotransmitter in controlling the release of red pigment-dispersing hormone', *Comparative and General Pharmacology*, 5: 299-303.
- Fitch, Walter M. 1970. 'Distinguishing homologous from analogous proteins', *Systematic zoology*, 19: 99-113.
- Forman, Henry Jay, Hongqiao Zhang, and Alessandra Rinna. 2009. 'Glutathione: overview of its protective roles, measurement, and biosynthesis', *Molecular aspects of medicine*, 30: 1-12.
- Fuller, Carl W, Lyle R Middendorf, Steven A Benner, George M Church, Timothy Harris, Xiaohua Huang, Stevan B Jovanovich, John R Nelson, Jeffery A Schloss, and David C Schwartz. 2009. 'The challenges of sequencing by synthesis', *Nature biotechnology*, 27: 1013-1023.
- Galloway, Tamara S, Rebecca J Brown, Mark A Browne, Awantha Dissanayake, David Lowe, Malcolm B Jones, and Michael H Depledge. 2004. 'A multibiomarker approach to environmental assessment', *Environmental Science & Technology*, 38: 1723-1731.
- Gertz, E Michael, Yi-Kuo Yu, Richa Agarwala, Alejandro A Schäffer, and Stephen F Altschul. 2006. 'Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST', *BMC biology*, 4: 1-14.
- Gilbert, D. 2010. 'EvidentialGene: Evidence Directed gene predictions for eukaryotes.' Available online at <http://arthropods.eugen.es/EvidentialGene/>.
- Gilbert, D. G. 2019a. 'Genes of the pig, *Sus scrofa*, reconstructed with EvidentialGene', *PeerJ*, 7.
- Gilbert, Don. 2016a. 'Perfect arthropod genes constructed with gigabases of RNA', *F1000Research*, 5.
- Gilbert, Don G. 2016b. 'Gene-omes built from mRNA seq not genome DNA', *F1000Research*, 5.

- Gilbert, Donald G. 2019b. 'Longest protein, longest transcript or most expression, for accurate gene reconstruction of transcriptomes?', *BioRxiv*: 829184.
- Goncalves, Angela, Andrew Tikhonov, Alvis Brazma, and Misha Kapushesky. 2011. 'A pipeline for RNA-seq data processing and quality assessment', *Bioinformatics*, 27: 867-869.
- Goode, George Brown. 1884. 'The fisheries and fishery industries of the United States', US Government Printing Office.
- Goodwin, Sara, John D McPherson, and W Richard McCombie. 2016. 'Coming of age: ten years of next-generation sequencing technologies', *Nature Reviews Genetics*, 17: 333-351.
- Grabherr, Manfred G, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, and Qiandong Zeng. 2011. 'Full-length transcriptome assembly from RNA-Seq data without a reference genome', *Nature biotechnology*, 29: 644-652.
- Grant, Gregory R, Elisabetta Manduchi, and Christian J Stoeckert Jr. 2007. 'Analysis and management of microarray gene expression data', *Current protocols in molecular biology*, 77: 19.6. 1-19.6. 30.
- Gravato, Carlos, Laura Guimarães, Joana Santos, Melissa Faria, Anabela Alves, and Lúcia Guilhermino. 2010. 'Comparative study about the effects of pollution on glass and yellow eels (*Anguilla anguilla*) from the estuaries of Minho, Lima and Douro Rivers (NW Portugal)', *Ecotoxicology and Environmental Safety*, 73: 524-533.
- Grosholz, Edwin D, and Gregory M Ruiz. 1996. 'Predicting the impact of introduced marine species: lessons from the multiple invasions of the European green crab *Carcinus maenas*', *Biological Conservation*, 78: 59-66.
- Guimaraes, Laura, Carlos Gravato, Joana Santos, Luís S Monteiro, and Lúcia Guilhermino. 2009. 'Yellow eel (*Anguilla anguilla*) development in NW Portuguese estuaries with different contamination levels', *Ecotoxicology*, 18: 385-402.
- Guimarães, Laura, Matias H Medina, and Lúcia Guilhermino. 2012. 'Health status of *Pomatoschistus microps* populations in relation to pollution and natural stressors: implications for ecological risk assessment', *Biomarkers*, 17: 62-77.
- Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. 2013. 'QUAST: quality assessment tool for genome assemblies', *Bioinformatics*, 29: 1072-1075.
- Guttman, Mitchell, Manuel Garber, Joshua Z Levin, Julie Donaghey, James Robinson, Xian Adiconis, Lin Fan, Magdalena J Koziol, Andreas Gnirke, and Chad Nusbaum. 2010. 'Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs', *Nature biotechnology*, 28: 503-510.
- Haas, Brian J, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, and Matthias Lieber. 2013. 'De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis', *Nature protocols*, 8: 1494-1512.
- Haas, Brian J, and Michael C Zody. 2010. 'Advancing RNA-seq analysis', *Nature biotechnology*, 28: 421-423.
- Hajibabaei, Mehrdad, Shadi Shokralla, Xin Zhou, Gregory AC Singer, and Donald J Baird. 2011. 'Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos', *PLoS One*, 6.
- Halim, Zahid. 2020. 'Optimizing the DNA fragment assembly using metaheuristic-based overlap layout consensus approach', *Applied Soft Computing*, 92: 106256.
- Harismendy, Olivier, Pauline C Ng, Robert L Strausberg, Xiaoyun Wang, Timothy B Stockwell, Karen Y Beeson, Nicholas J Schork, Sarah S Murray, Eric J Topol, and Samuel Levy. 2009. 'Evaluation of next generation sequencing platforms for population targeted sequencing studies', *Genome biology*, 10: 1-13.
- Harrison, Peter W, Alison E Wright, and Judith E Mank. 2012. 'The evolution of gene expression and the transcriptome phenotype relationship', *Seminars in cell & developmental biology*, 222-229.

- He, Yiming, Zhen Zhang, Xiaoqing Peng, Fangxiang Wu, and Jianxin Wang. 2013. 'De novo assembly methods for next generation sequencing data', *Tsinghua Science and Technology*, 18: 500-514.
- Henschel, Robert, Matthias Lieber, Le-Shin Wu, Phillip M Nista, Brian J Haas, and Richard D LeDuc. 2012. 'Trinity RNA-Seq assembler performance optimization', *Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the extreme to the campus and beyond*, 1-8.
- Hillier, LaDeana W, Gabor T Marth, Aaron R Quinlan, David Dooling, Ginger Fewell, Derek Barnett, Paul Fox, Jarret I Glasscock, Matthew Hickenbotham, and Weichun Huang. 2008. 'Whole-genome sequencing and variant discovery in *C. elegans*', *Nature methods*, 5: 183-188.
- Hong, Mingye, Shuang Tao, Ling Zhang, Li-Ting Diao, Xuanmei Huang, Shaohui Huang, Shu-Juan Xie, Zhen-Dong Xiao, and Hua Zhang. 2020. 'RNA sequencing: new technologies and applications in cancer research', *Hematology & oncology*, 13: 1-16.
- Hopkins, Penny M. 2012. 'The eyes have it: A brief history of crustacean neuroendocrinology', *General and comparative endocrinology*, 175: 357-366.
- Huang, Xin, Xiao-Guang Chen, and Peter A Armbruster. 2016. 'Comparative performance of transcriptome assembly methods for non-model organisms', *BMC genomics*, 17: 1-14.
- Hudson, Matthew E. 2008. 'Sequencing breakthroughs for genomic ecology and evolutionary biology', *Molecular ecology resources*, 8: 3-17.
- Hutchison III, Clyde A. 2007. 'DNA sequencing: bench to bedside and beyond', *Nucleic acids research*, 35: 6227-6237.
- Hyatt, Doug, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. 2010. 'Prodigal: prokaryotic gene recognition and translation initiation site identification', *BMC bioinformatics*, 11: 1-11.
- Idury, Ramana M, and Michael S Waterman. 1995. 'A new algorithm for DNA sequence assembly', *Computational biology*, 2: 291-306.
- Jauhal, April A, and Richard D Newcomb. 2021. 'Assessing genome assembly quality prior to downstream analysis: N50 versus BUSCO', *Molecular Ecology Resources*, 21: 1416-1421.
- Jolliffe, Ian T, and Jorge Cadima. 2016. 'Principal component analysis: a review and recent developments', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374: 20150202.
- Kaiser, Michel J, Roger N Hughes, and David G Reid. 1990. 'Chelal morphometry, prey-size selection and aggressive competition in green and red forms of *Carcinus maenas* (L.)', *Experimental marine biology and ecology*, 140: 121-134.
- Kegel, Günter, Barbara Reichwein, Cornelis P Tensen, and Rainer Keller. 1991. 'Amino acid sequence of crustacean hyperglycemic hormone (CHH) from the crayfish, *Orconectes limosus*: Emergence of a novel neuropeptide family', *Peptides*, 12: 909-913.
- Kegel, Günter, Barbara Reichwein, Susanne Weese, Gabriele Gaus, Jasna Peter-Katalinić, and Rainer Keller. 1989. 'Amino acid sequence of the crustacean hyperglycemic hormone (CHH) from the shore crab, *Carcinus maenas*', *FEBS letters*, 255: 10-14.
- Keller, Oliver, Martin Kollmar, Mario Stanke, and Stephan Waack. 2011. 'A novel hybrid gene prediction method employing protein multiple sequence alignments', *Bioinformatics*, 27: 757-763.
- Kircher, Martin, Patricia Heyn, and Janet Kelso. 2011. 'Addressing challenges in the production and analysis of illumina sequencing data', *BMC genomics*, 12: 1-14.
- Kircher, Martin, and Janet Kelso. 2010. 'High-throughput DNA sequencing—concepts and limitations', *Bioessays*, 32: 524-536.

- Klassen, Gregory John, and Andrea Locke. 2007. A biological synopsis of the European green crab, *Carcinus maenas*, pp. 1-75.
- Knigge, Thomas, Gerald A LeBlanc, and Alex T Ford. 2021. 'A crab is not a fish: unique aspects of the crustacean endocrine system and considerations for endocrine toxicology', *Frontiers in Endocrinology*, 12: 98.
- Koonin, Eugene V. 2005. 'Orthologs, paralogs, and evolutionary genomics', *Annual review of genetics*, 39: 309-338.
- Kukurba, Kimberly R, and Stephen B Montgomery. 2015. 'RNA sequencing and analysis', *Cold Spring Harbor Protocols*, 951-969.
- Lall, Snehalika, Debajyoti Sinha, Sanghamitra Bandyopadhyay, and Debarka Sengupta. 2018. 'Structure-aware principal component analysis for single-cell RNA-seq data', *Computational Biology*, 25: 1365-1373.
- Langmead, Ben, and Steven L Salzberg. 2012. 'Fast gapped-read alignment with Bowtie 2', *Nature methods*, 9: 357-359.
- Laufer, Hans, William J Biggers, and Jonna SB Ahl. 1998. 'Stimulation of ovarian maturation in the crayfish *Procambarus clarkii* by methyl farnesoate', *General and comparative endocrinology*, 111: 113-118.
- Ledesma, Federico M, Silvina Van der Molen, and Pedro J Barón. 2010. 'Sex identification of *Carcinus maenas* by analysis of carapace geometrical morphometry', *Sea Research*, 63: 213-216.
- Lee, Heewook, and Haixu Tang. 2012. 'Next-generation sequencing technologies and fragment assembly algorithms', *Evolutionary genomics*: 155-174.
- Leignel, VSJH, JH Stillman, S Baringou, R Thabet, and I Metais. 2014. 'Overview on the European green crab *Carcinus* spp.(Portunidae, Decapoda), one of the most famous marine invaders and ecotoxicological models', *Environmental Science and Pollution Research*, 21: 9129-9144.
- Leuven, RSEW, PP Jaros, F Van Herp, and R Keller. 1982. 'Species or group specificity in biological and immunological studies of crustacean hyperglycemic hormone', *General and comparative endocrinology*, 46: 288-296.
- Levy Karin, Eli, Milot Mirdita, and Johannes Söding. 2020. 'MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics', *Microbiome*, 8: 1-15.
- Li, Ruiqiang, Chang Yu, Yingrui Li, Tak-Wah Lam, Siu-Ming Yiu, Karsten Kristiansen, and Jun Wang. 2009. 'SOAP2: an improved ultrafast tool for short read alignment', *Bioinformatics*, 25: 1966-1967.
- Li, Ruiqiang, Hongmei Zhu, Jue Ruan, Wubin Qian, Xiaodong Fang, Zhongbin Shi, Yingrui Li, Shengting Li, Gao Shan, and Karsten Kristiansen. 2010. 'De novo assembly of human genomes with massively parallel short read sequencing', *Genome research*, 20: 265-272.
- Li, Weizhong, and Adam Godzik. 2006. 'Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences', *Bioinformatics*, 22: 1658-1659.
- Li, Weizhong, Lukasz Jaroszewski, and Adam Godzik. 2001. 'Clustering of highly homologous sequences to reduce the size of large protein databases', *Bioinformatics*, 17: 282-283.
- Lischer, Heidi EL, and Kentaro K Shimizu. 2017. 'Reference-guided de novo assembly approach improves genome reconstruction for related species', *BMC bioinformatics*, 18: 1-12.
- Liu, Lin, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, and Maggie Law. 2012. 'Comparison of next-generation sequencing systems', *J Biomed Biotechnol*, 2012: 251364.
- Lo, Chien-Chi, and Patrick SG Chain. 2014. 'Rapid evaluation and quality control of next generation sequencing data with FaQCs', *BMC bioinformatics*, 15: 1-8.

- Lu, BingXin, ZhenBing Zeng, and TieLiu Shi. 2013. 'Comparative study of de novo assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq', *Science China Life Sciences*, 56: 143-155.
- Luo, Ruibang, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, Yanxiang Chen, Qi Pan, and Yunjie Liu. 2012. 'SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler', *Gigascience*, 1: 2047-217X-1-18.
- Maher, Christopher A, Chandan Kumar-Sinha, Xuhong Cao, Shanker Kalyana-Sundaram, Bo Han, Xiaojun Jing, Lee Sam, Terrence Barrette, Nallasivam Palanisamy, and Arul M Chinnaiyan. 2009. 'Transcriptome sequencing to detect gene fusions in cancer', *Nature*, 458: 97-101.
- Manni, Mosè, Matthew R Berkeley, Mathieu Seppey, Felipe A Simão, and Evgeny M Zdobnov. 2021. 'BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes', *Molecular biology and evolution*, 38: 4647-4654.
- Manni, Mosè, Matthew R Berkeley, Mathieu Seppey, and Evgeny M Zdobnov. 2021. 'BUSCO: assessing genomic data quality and beyond', *Current Protocols*, 1.
- Marçais, Guillaume, and Carl Kingsford. 2011. 'A fast, lock-free approach for efficient parallel counting of occurrences of k-mers', *Bioinformatics*, 27: 764-770.
- Mardis, Elaine R. 2008. 'Next-generation DNA sequencing methods', *Annual review of genomics and human genetics*, 9: 387-402.
- Margulies, Marcel, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bembien, Jan Berka, Michael S Braverman, Yi-Ju Chen, and Zhoutao Chen. 2005. 'Genome sequencing in microfabricated high-density picolitre reactors', *Nature*, 437: 376-380.
- Martin, Jeffrey A, and Zhong Wang. 2011. 'Next-generation transcriptome assembly', *Nature Reviews Genetics*, 12: 671-682.
- Martins, Irene, Ester Dias, Martina I Ilarri, Francisco J Campuzano, Lúgia Pinto, Miguel M Santos, and Carlos Antunes. 2019. 'Antagonistic effects of multiple stressors on macroinvertebrate biomass from a temperate estuary (Minho estuary, NW Iberian Peninsula)', *Ecological Indicators*, 101: 792-803.
- McCarthy, Alice. 2010. 'Third generation DNA sequencing: pacific biosciences' single molecule real time technology', *Chemistry & biology*, 17: 675-676.
- McDermaid, Adam, Brandon Monier, Jing Zhao, Bingqiang Liu, and Qin Ma. 2019. 'Interpretation of differential gene expression results of RNA-seq data: review and integration', *Briefings in bioinformatics*, 20: 2044-2054.
- McGaw, Iain J, MJ Kaiser, E Naylor, and Roger N Hughes. 1992. 'Intraspecific morphological variation related to the moult-cycle in colour forms of the shore crab *Carcinus maenas*', *Zoology*, 228: 351-359.
- Meire, Patrick, Tom Ysebaert, Stefan Van Damme, Erika Van den Bergh, Tom Maris, and Eric Struyf. 2005. 'The Scheldt estuary: a description of a changing ecosystem', *Hydrobiologia*, 540: 1-11.
- Meng, Yanling, and Enmin Zou. 2009. 'A molecular biomarker for disruption of crustacean molting: the N-acetyl- β -glucosaminidase mRNA in the epidermis of the fiddler crab', *Bulletin of environmental contamination and toxicology*, 82: 554-558.
- Mesquita, Sofia Raquel, Lúcia Guilhermino, and Laura Guimarães. 2011. 'Biochemical and locomotor responses of *Carcinus maenas* exposed to the serotonin reuptake inhibitor fluoxetine', *Chemosphere*, 85: 967-976.
- Metzker, Michael L. 2005. 'Emerging technologies in DNA sequencing', *Genome research*, 15: 1767-1776.
- Metzker, Michael L. 2010. 'Sequencing technologies—the next generation', *Nature reviews genetics*, 11: 31-46.

- Mil-Homens, Mário, AM Costa, S Fonseca, Maria Ascensão Trancoso, C Lopes, R Serrano, and R Sousa. 2013a. 'Characterization of heavy-metal contamination in surface sediments of the Minho River Estuary by way of factor analysis', *Archives of environmental contamination and toxicology*, 64: 617-631.
- Mil-Homens, Mário, AM Costa, S Fonseca, Maria Ascensão Trancoso, C Lopes, R Serrano, and R Sousa. 2013b. 'Natural heavy metal and metalloid concentrations in sediments of the Minho River estuary (Portugal): baseline values for environmental studies', *Environmental monitoring and assessment*, 185: 5937-5950.
- Monteiro, M, Carla Quintaneiro, AJA Nogueira, F Morgado, AMVM Soares, and L Guilhermino. 2007. 'Impact of chemical exposure on the fish *Pomatoschistus microps* Krøyer (1838) in estuaries of the Portuguese Northwest coast', *Chemosphere*, 66: 514-522.
- Montgomery, Stephen B, Micha Sammeth, Maria Gutierrez-Arcelus, Radoslaw P Lach, Catherine Ingle, James Nisbett, Roderic Guigo, and Emmanouil T Dermizakis. 2010. 'Transcriptome genetics using second generation sequencing in a Caucasian population', *Nature*, 464: 773-777.
- Moss, Dorian, and BK Wyatt. 1994. 'The CORINE biotopes project: a database for conservation of nature and wildlife in the European community', *Applied Geography*, 14: 327-349.
- Mota, Micaela, Ana Bio, Miguel Bao, Santiago Pascual, Eric Rochard, and Carlos Antunes. 2015. 'New insights into biology and ecology of the Minho River Allis shad (*Alosa alosa* L.): contribution to the conservation of one of the last European shad populations', *Reviews in Fish Biology and Fisheries*, 25: 395-412.
- Mykles, Donald L. 2011. 'Ecdysteroid metabolism in crustaceans', *Steroid biochemistry and molecular biology*, 127: 196-203.
- Mykles, Donald L, and Ernest S Chang. 2020. 'Hormonal control of the crustacean molting gland: Insights from transcriptomics and proteomics', *General and comparative endocrinology*, 294: 113493.
- Nakasugi, Kenlee, Ross Crowhurst, Julia Bally, and Peter Waterhouse. 2014. 'Combining Transcriptome Assemblies from Multiple De Novo Assemblers in the Allo-Tetraploid Plant *Nicotiana benthamiana*', *PLoS One*, 9.
- Nayfach, Stephen, Antonio Pedro Camargo, Frederik Schulz, Emiley Eloie-Fadrosch, Simon Roux, and Nikos C Kyrpides. 2021. 'CheckV assesses the quality and completeness of metagenome-assembled viral genomes', *Nature biotechnology*, 39: 578-585.
- Nielsen, Torsten O, Rob B West, Sabine C Linn, Orly Alter, Margaret A Knowling, John X O'Connell, Shirley Zhu, Mike Fero, Gavin Sherlock, and Jonathan R Pollack. 2002. 'Molecular characterisation of soft tissue tumours: a gene expression study', *The Lancet*, 359: 1301-1307.
- Nong, Wenyan, Zacary YH Chai, Xiaosen Jiang, Jing Qin, Ka Yan Ma, King Ming Chan, Ting Fung Chan, Billy KC Chow, Hoi Shan Kwan, and Chris KC Wong. 2020. 'A crustacean annotated transcriptome (CAT) database', *BMC genomics*, 21: 1-5.
- Nyrén, Pettersson, Bertil Pettersson, and Mathias Uhlén. 1993. 'Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay', *Analytical biochemistry*, 208: 171-175.
- Oberman, Laurence, and Mark Seger. 2007. 'Collectl'. <http://collectl.sourceforge.net/index.html>.
- Ono, Hanako, Kazuo Ishii, Toshinori Kozaki, Isao Ogiwara, Motoki Kanekatsu, and Tetsuya Yamada. 2015. 'Removal of redundant contigs from de novo RNA-Seq assemblies via homology search improves accurate detection of differentially expressed genes', *BMC genomics*, 16: 1-13.
- Oshlack, Alicia, Mark D Robinson, and Matthew D Young. 2010. 'From RNA-seq reads to differential expression results', *Genome Biology*, 11: 1-10.

- Ozsolak, Fatih, and Patrice M Milos. 2011. 'RNA sequencing: advances, challenges and opportunities', *Nature Reviews Genetics*, 12: 87-98.
- Pareek, Chandra Shekhar, Rafal Smoczynski, and Andrzej Tretyn. 2011. 'Sequencing technologies and genome sequencing', *Applied genetics*, 52: 413-435.
- Parks, Donovan H, Michael Imelfort, Connor T Skennerton, Philip Hugenholtz, and Gene W Tyson. 2015. 'CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes', *Genome research*, 25: 1043-1055.
- Parra, Genis, Keith Bradnam, and Ian Korf. 2007. 'CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes', *Bioinformatics*, 23: 1061-1067.
- Patel, Ravi K, and Mukesh Jain. 2012. 'NGS QC Toolkit: a toolkit for quality control of next generation sequencing data', *PLoS One*, 7.
- Patro, Rob, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. 2017. 'Salmon provides fast and bias-aware quantification of transcript expression', *Nature methods*, 14: 417-419.
- Pedersen, KL, SN Pedersen, P Højrup, Jens S Andersen, P Roepstorff, Jens Knudsen, and MH Depledge. 1994. 'Purification and characterization of a cadmium-induced metallothionein from the shore crab *Carcinus maenas* (L.)', *Biochemical*, 297: 609-614.
- Peng, Jinrong. 2019. 'Gene redundancy and gene compensation: An updated view', *Genetics and genomics*, 46: 329-333.
- Pereira, Patrícia, Hilda de Pablo, Maria Dulce Subida, Carlos Vale, and Mário Pacheco. 2011. 'Bioaccumulation and biochemical markers in feral crab (*Carcinus maenas*) exposed to moderate environmental contamination—The impact of non-contamination-related variables', *Environmental toxicology*, 26: 524-540.
- Pertea, Mihaela, Daehwan Kim, Geo M Pertea, Jeffrey T Leek, and Steven L Salzberg. 2016. 'Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown', *Nature Protocols*, 11: 1650-1667.
- Pfrender, MichaelE, CharlesP Hawkins, Mark Bagley, GregoryW Courtney, BrianR Creutzburg, JohnH Epler, Steve Fend, LeonardC Ferrington Jr, PaulaL Hartzell, and Suzanne Jackson. 2010. 'Assessing macroinvertebrate biodiversity in freshwater ecosystems: advances and challenges in DNA-based approaches', *The Quarterly review of biology*, 85: 319-340.
- Picado, Ana, Maria João Bebianno, MH Costa, A Ferreira, and C Vale. 2007. 'Biomarkers: a strategic tool in the assessment of environmental quality of coastal waters', *Hydrobiologia*, 587: 79-87.
- Pinto, Rute, and Filomena Cardoso Martins. 2013. 'The Portuguese national strategy for integrated coastal zone management as a spatial planning instrument to climate change adaptation in the Minho River Estuary (Portugal NW-Coastal Zone)', *Environmental science & policy*, 33: 76-96.
- Pop, Mihai, and Steven L Salzberg. 2008. 'Bioinformatics challenges of new sequencing technology', *Trends in genetics*, 24: 142-149.
- Priedhorsky, Reid, and Tim Randles. 2017. 'Charliecloud: Unprivileged containers for user-defined software stacks in hpc', *Proceedings of the international conference for high performance computing, networking, storage and analysis*, 1-10.
- Queiroga, Henrique, John D Costlow, and Maria Helena Moreira. 1997. 'Vertical migration of the crab *Carcinus maenas* first zoea in an estuary: implications for tidal stream transport', *Marine Ecology Progress Series*, 149: 121-132.
- Ramos, Sandra, Robert K Cowen, Claire Paris, Pedro Ré, and Adriano A Bordalo. 2006. 'Environmental forcing and larval fish assemblage dynamics in the Lima River estuary (northwest Portugal)', *Plankton Research*, 28: 275-286.

- Ray, Gary L. 2005. 'Invasive animal species in marine and estuarine environments: Biology and ecology'. U.S. Army Corp of Engineers, U.S. Army Engineer Research and Development Center Environmental Laboratory. Pp 64.
- Reid, DG, P Abelló, MJ Kaiser, and CG Warman. 1997. 'Carapace colour, inter-moult duration and the behavioural and physiological ecology of the shore crab *Carcinus maenas*', Estuarine, Coastal and Shelf Science, 44: 203-211.
- Reid, DG, and JC Aldrich. 1989. 'Variations in response to environmental hypoxia of different colour forms of the shore crab, *Carcinus maenas*', Comparative Biochemistry and Physiology Part A: Physiology, 92: 535-539.
- Reuter, Jason A, Damek V Spacek, and Michael P Snyder. 2015. 'High-throughput sequencing technologies', Molecular cell, 58: 586-597.
- Ringnér, Markus. 2008. 'What is principal component analysis?', Nature Biotechnology, 26: 303-304.
- Rio, Donald C, Manuel Ares, Gregory J Hannon, and Timothy W Nilsen. 2010. 'Purification of RNA using TRIzol (TRI reagent)', Cold Spring Harbor Protocols, 6.
- Rodrigues, AP, C Gravato, and L Guimaraes. 2013. 'Involvement of the antioxidant system in differential sensitivity of *Carcinus maenas* to fenitrothion exposure', Environmental Science: Processes & Impacts, 15: 1938-1948.
- Rodrigues, Aurélie P, Lúcia HMLM Santos, Maria Teresa Oliva-Teles, Cristina Delerue-Matos, and Laura Guimarães. 2014. 'Joint effects of salinity and the antidepressant sertraline on the estuarine decapod *Carcinus maenas*', Aquatic toxicology, 156: 169-178.
- Rodrigues, Aurélie P, Lúcia HMLM Santos, Maria João Ramalhosa, Cristina Delerue-Matos, and Laura Guimarães. 2015. 'Sertraline accumulation and effects in the estuarine decapod *Carcinus maenas*: importance of the history of exposure to chemical stress', Hazardous Materials, 283: 350-358.
- Rodrigues, Aurélie Pinto, Teresa Oliva-Teles, Sofia Raquel Mesquita, Cristina Delerue-Matos, and Laura Guimarães. 2014. 'Integrated biomarker responses of an estuarine invertebrate to high abiotic stress and decreased metal contamination', Marine environmental research, 101: 101-114.
- Rodrigues, Aurélie Pinto, Patrícia Correia Oliveira, Lúcia Guilhermino, and Laura Guimaraes. 2012. 'Effects of salinity stress on neurotransmission, energy metabolism, and anti-oxidant biomarkers of *Carcinus maenas* from two estuaries of the NW Iberian Peninsula', Marine biology, 159: 2061-2074.
- Rodrigues, Elsa Teresa, and Miguel Ângelo Pardal. 2014. 'The crab *Carcinus maenas* as a suitable experimental model in ecotoxicology', Environment international, 70: 158-182.
- Rodríguez, Enrique M, Daniel A Medesani, and Milton Fingerman. 2007. 'Endocrine disruption in crustaceans due to pollutants: a review', Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology, 146: 661-671.
- Roer, Robert, and Richard Dillaman. 1984. 'The structure and calcification of the crustacean cuticle', American Zoologist, 24: 893-909.
- Ronaghi, Mostafa, Samer Karamohamed, Bertil Pettersson, Mathias Uhlén, and Pål Nyrén. 1996. 'Real-time DNA sequencing using detection of pyrophosphate release', Analytical biochemistry, 242: 84-89.
- ROPEZ, JW. 1968. 'The feeding habits of the green crab, *Carcinus maenas* (L.)', Fisheries Bulletin, 67: 183-203.
- Saary, Paul, Alex L Mitchell, and Robert D Finn. 2020. 'Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC', Genome biology, 21: 1-21.
- Sanger, Frederick, Steven Nicklen, and Alan R Coulson. 1977. 'DNA sequencing with chain-terminating inhibitors', Proceedings of the national academy of sciences, 74: 5463-5467.

- Santos, Euclides Antonio, Luiz Eduardo, Maia Nery, Alex Augusto Gonçalves, and Rainer Keller. 1997. 'Evidence for the involvement of the crustacean hyperglycemic hormone in the regulation of lipid metabolism', *Physiological Zoology*, 70: 415-420.
- Saraiva, Sofia, P Pina, F Martins, M Santos, F Braunschweig, and R Neves. 2007. 'Modelling the influence of nutrient loads on Portuguese estuaries', *Hydrobiologia*, 587: 5-18.
- Schena, Mark, Dari Shalon, Ronald W Davis, and Patrick O Brown. 1995. 'Quantitative monitoring of gene expression patterns with a complementary DNA microarray', *Science*, 270: 467-470.
- Schulz, Marcel H, Daniel R Zerbino, Martin Vingron, and Ewan Birney. 2012. 'Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels', *Bioinformatics*, 28: 1086-1092.
- Sedlmeier, Dietrich. 1985. 'Mode of action of the crustacean hyperglycemic hormone', *American Zoologist*, 25: 223-232.
- Sedlmeier, Dietrich. 1988. 'The crustacean hyperglycemic hormone (CHH) releases amylase from the crayfish midgut gland', *Regulatory peptides*, 20: 91-98.
- Seppely, Mathieu, Mosè Manni, and Evgeny M Zdobnov. 2019. 'BUSCO: assessing genome assembly and annotation completeness.' *Gene prediction*, p. 227-245.
- Waterson, Robert H., Eric S. Lander, and Richard K. Wilson. 2005. 'Initial sequence of the chimpanzee genome and comparison with the human genome', *Nature*, 437: 69-87.
- Sheharyar, Ali, Talar Boghos Yacoubian, Dina Aljogol, Borbala Mifsud, Dena Al Thani, and Michael Aupetit. 2020. 'An enhanced MA plot with R-Shiny to ease exploratory analysis of transcriptomic data', *ArXiv*.
- Shendure, Jay. 2008. 'The beginning of the end for microarrays?', *Nature methods*, 5: 585-587.
- Shendure, Jay, Gregory J Porreca, Nikos B Reppas, Xiaoxia Lin, John P McCutcheon, Abraham M Rosenbaum, Michael D Wang, Kun Zhang, Robi D Mitra, and George M Church. 2005. 'Accurate multiplex polony sequencing of an evolved bacterial genome', *Science*, 309: 1728-1732.
- Shokralla, Shadi, Jennifer L Spall, Joel F Gibson, and Mehrdad Hajibabaei. 2012. 'Next-generation sequencing technologies for environmental DNA research', *Molecular ecology*, 21: 1794-1805.
- Shrestha, Anish Man Singh, and Martin C Frith. 2013. 'An approximate Bayesian approach for mapping paired-end DNA reads to a reference genome', *Bioinformatics*, 29: 965-972.
- Sieber, Patricia, Matthias Platzter, and Stefan Schuster. 2018. 'The definition of open reading frame revisited', *Trends in genetics*, 34: 167-170.
- Simão, Felipe A, Robert M Waterhouse, Panagiotis Ioannidis, Evgenia V Kriventseva, and Evgeny M Zdobnov. 2015. 'BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs', *Bioinformatics*, 31: 3210-3212.
- Simmers, AJ, and BMH Bush. 1983. 'Central nervous mechanisms controlling rhythmic burst generation in the ventilatory motoneurons of *Carcinus maenas*', *Comparative physiology*, 150: 1-21.
- Lander, Eric S. 2002. 'Initial sequencing and comparative analysis of the mouse genome', *Nature*, 420: 520-562.
- Smith-Unna, Richard, Chris Boursnell, Rob Patro, Julian M Hibberd, and Steven Kelly. 2016. 'TransRate: reference-free quality assessment of de novo transcriptome assemblies', *Genome research*, 26: 1134-1144.
- Sohn, Jang-il, and Jin-Wu Nam. 2018. 'The present and future of de novo whole-genome assembly', *Briefings in bioinformatics*, 19: 23-40.

- Sorenson, AL. 1973. 'Demonstration of an action of acetylcholine on the central nervous system of a crab', *The Biological Bulletin*, 144: 180-191.
- Sousa, Ronaldo, Sérgio Dias, and Carlos Antunes. 2007. "Subtidal macrobenthic structure in the lower Lima estuary, NW of Iberian Peninsula." *Annales Zoologici Fennici*, p. 303-313.
- Sousa, Ronaldo, Sérgio Dias, Vânia Freitas, and Carlos Antunes. 2008. 'Subtidal macrozoobenthic assemblages along the River Minho estuarine gradient (north-west Iberian Peninsula)', *Aquatic Conservation*, 18: 1063-1077.
- Sousa, Ronaldo Gomes de. 2003. 'Estruturas das comunidades de macroinvertebrados bentônicos presentes no estuário do Rio Lima', [Master's thesis, Instituto de Ciências Biomédicas Abel Salazar do Porto], <http://hdl.handle.net/10216/9772>.
- Sousa, Ronaldo, Lúcia Guilhermino, and Carlos Antunes. 2005. 'Molluscan fauna in the freshwater tidal area of the River Minho estuary, NW of Iberian Peninsula', *Annales de Limnologie-International Journal of Limnology*, p. 141-147.
- Souza, Allan T, Ester Dias, Ana Nogueira, Joana Campos, João C Marques, and Irene Martins. 2013. 'Population ecology and habitat preferences of juvenile flounder *Platichthys flesus* (Actinopterygii: Pleuronectidae) in a temperate estuary', *Sea Research*, 79: 60-69.
- Soyez, D, JP Le Caer, PY Noel, and J Rossier. 1991. 'Primary structure of two isoforms of the vitellogenesis inhibiting hormone from the lobster *Homarus americanus*', *Neuropeptides*, 20: 25-32.
- Spindler, Klaus-Dieter, Dieter Adelung, and Catherine Tchernigovtzeff. 1974. 'A comparison of the methods of molt staging according to Drach and to Adelung in the common shore crab, *Carcinus maenas*', *Zeitschrift für Naturforschung C*, 29: 754-756.
- Stanke, Mario, and Burkhard Morgenstern. 2005. 'AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints', *Nucleic acids research*, 33: 465-467.
- Stark, Rory, Marta Grzelak, and James Hadfield. 2019. 'RNA sequencing: the teenage years', *Nature Reviews Genetics*, 20: 631-656.
- Sun, Liang, Sufen Dong, Yinbing Ge, Jose Pedro Fonseca, Zachary T Robinson, Kirankumar S Mysore, and Perdeep Mehta. 2019. 'DiVenn: an interactive and integrated web-based visualization tool for comparing gene lists', *Frontiers in genetics*: 421.
- Swerdlow, Harold, and Raymond Gesteland. 1990. 'Capillary gel electrophoresis for rapid, high resolution DNA sequencing', *Nucleic acids research*, 18: 1415-1419.
- Trapnell, Cole, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J Van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. 2010. 'Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation', *Nature biotechnology*, 28: 511-515.
- Travis, D. F. 1965. 'The deposition of skeletal structures in the crustacea. 5. The histomorphological and histochemical changes associated with the development and calcification of the branchial exoskeleton in the crayfish, *Orconectes virilis* Hagen', *Acta Histochem*, 20: 193-222.
- Travis, Dorothy F. 1955. 'The molting cycle of the spiny lobster, *Panulirus argus* Latreille. III. Physiological changes which occur in the blood and urine during the normal molting cycle', *The Biological Bulletin*, 109: 484-503.
- Treerattrakool, Supattra, Sakol Panyim, Siu-Ming Chan, Boonsirm Withyachumnarnkul, and Apinunt Udomkit. 2008. 'Molecular characterization of gonad-inhibiting hormone of *Penaeus monodon* and elucidation of its inhibitory role in vitellogenin expression by RNA interference', *The Febs*, 275: 970-980.

- Trivedi, Urmi H, Timothée Cézard, Stephen Bridgett, Anna Montazam, Jenna Nichols, Mark Blaxter, and Karim Gharbi. 2014. 'Quality control of next-generation sequencing data without a reference', *Frontiers in genetics*, 5: 111.
- Van Rossum, Guido, and Fred L Drake Jr. 1995. 'Python tutorial', Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 620.
- Verbruggen, Bas, Lisa K Bickley, Eduarda M Santos, Charles R Tyler, Grant D Stentiford, Kelly S Bateman, and Ronny van Aerle. 2015. 'De novo assembly of the *Carcinus maenas* transcriptome and characterization of innate immune system pathways', *BMC genomics*, 16: 1-17.
- Vezzi, Francesco, Giuseppe Narzisi, and Bud Mishra. 2012. 'Feature-by-feature—evaluating de novo sequence assembly', *PloS One*, 7.
- Vieira, Luís R, Carlos Gravato, AMVM Soares, F Morgado, and L Guilhermino. 2009. 'Acute effects of copper and mercury on the estuarine fish *Pomatoschistus microps*: linking biomarkers to behaviour', *Chemosphere*, 76: 1416-1427.
- Voshall, Adam, and Etsuko N Moriyama. 2018. 'Next-generation transcriptome assembly: strategies and performance analysis', *Bioinformatics in the era of post genomics and big data*: 15-36.
- Wainwright, Geoffrey, Simon G Webster, Mark C Wilkinson, J Sook Chung, and Huw H Rees. 1996. 'Structure and significance of mandibular organ-inhibiting hormone in the crab, *Cancer pagurus*: involvement in multihormonal regulation of growth and reproduction', *Biological Chemistry*, 271: 12749-12754.
- Wajid, Bilal, and Erchin Serpedin. 2012. 'Review of general algorithmic features for genome assemblers for next generation sequencers', *Genomics, proteomics & bioinformatics*, 10: 58-73.
- Wajid, Bilal, and Erchin Serpedin. 2014. 'Do it yourself guide to genome assembly', *Briefings in Functional Genomics*, 15: 1-9.
- Wall, Jeffrey D, Ling Fung Tang, Brandon Zerbe, Mark N Kvale, Pui-Yan Kwok, Catherine Schaefer, and Neil Risch. 2014. 'Estimating genotype error rates from high-coverage next-generation sequence data', *Genome research*, 24: 1734-1739.
- Wang, Eric T, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F Kingsmore, Gary P Schroth, and Christopher B Burge. 2008. 'Alternative isoform regulation in human tissue transcriptomes', *Nature*, 456: 470-476.
- Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. 'RNA-Seq: a revolutionary tool for transcriptomics', *Nature reviews genetics*, 10: 57-63.
- Waterhouse, Robert M, Mathieu Seppey, Felipe A Simão, Mosè Manni, Panagiotis Ioannidis, Guennadi Klioutchnikov, Evgenia V Kriventseva, and Evgeny M Zdobnov. 2018. 'BUSCO applications from quality assessments to gene prediction and phylogenomics', *Molecular biology and evolution*, 35: 543-548.
- Waterhouse, Robert M, Mathieu Seppey, Felipe A Simão, and Evgeny M Zdobnov. 2019. 'Using BUSCO to assess insect genomic resources', *Insect genomics*, p. 59-74.
- Waterhouse, Robert M, Evgeny M Zdobnov, and Evgenia V Kriventseva. 2011. 'Correlating traits of gene retention, sequence divergence, duplicability and essentiality in vertebrates, arthropods, and fungi', *Genome biology and evolution*, 3: 75-86.
- Webster, SG. 1991. 'Amino acid sequence of putative moult-inhibiting hormone from the crab *Carcinus maenas*', *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 244: 247-252.

- Webster, Simon George, Rainer Keller, and Heinrich Dirksen. 2012. 'The CHH-superfamily of multifunctional peptide hormones controlling crustacean metabolism, osmoregulation, moulting, and reproduction', *General and comparative endocrinology*, 175: 217-233.
- Xuereb, Benoît, Estelle Lefèvre, Jeanne Garric, and Olivier Geffard. 2009. 'Acetylcholinesterase activity in *Gammarus fossarum* (Crustacea Amphipoda): linking AChE inhibition and behavioural alteration', *Aquatic toxicology*, 94: 114-122.
- Yang, Jian-Hua, Peng Shao, Hui Zhou, Yue-Qin Chen, and Liang-Hu Qu. 2010. 'deepBase: a database for deeply annotating and mining deep sequencing data', *Nucleic acids research*, 38: 123-130.
- Yang, Xi, Di Liu, Fei Liu, Jun Wu, Jing Zou, Xue Xiao, Fangqing Zhao, and Baoli Zhu. 2013. 'HTQC: a fast quality control toolkit for Illumina sequencing data', *BMC bioinformatics*, 14: 1-4.
- Ylonen, Tatu, and Chris Lonvick. 2006. 'The secure shell (SSH) protocol architecture.'
- Zdobnov, Evgeny M, Dmitry Kuznetsov, Fredrik Tegenfeldt, Mosè Manni, Matthew Berkeley, and Evgenia V Kriventseva. 2021. 'OrthoDB in 2020: evolutionary and functional annotations of orthologs', *Nucleic acids research*, 49: 389-393.
- Zeng, Chaoshu, and Ernest Naylor. 1996. 'Endogenous tidal rhythms of vertical migration in field collected zoea-1 larvae of the shore crab *Carcinus maenas*: implications for ebb tide offshore dispersal', *Marine Ecology Progress Series*, 132: 71-82.
- Zhang, Hui, Nathan Hall, Leslie R Goertzen, Charles Y Chen, Eric Peatman, Jinesh Patel, and J Scott McElroy. 2019. 'Transcriptome Analysis Reveals Unique Relationships Among Eleusine Species and Heritage of *Eleusine coracana*', *G3: Genes, Genomes, Genetics.*, 9: 2029-2036.
- Zhao, Qiong-Yi, Yi Wang, Yi-Meng Kong, Da Luo, Xuan Li, and Pei Hao. 2011. 'Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study', *BMC bioinformatics*, 1-12.
- Zhou, Qian, Xiaoquan Su, Gongchao Jing, Songlin Chen, and Kang Ning. 2018. 'RNA-QC-chain: comprehensive and fast quality control for RNA-Seq data', *BMC genomics*, 19: 1-10.
- Zhou, Qian, Xiaoquan Su, Anhui Wang, Jian Xu, and Kang Ning. 2013. 'QC-Chain: fast and holistic quality control method for next-generation sequencing data', *PloS One*, 8.