Universidade de São Paulo
Instituto de Matemática e Estatística
Programa Interunidades de Pós-Graduação em Bioinformática
(Mestrado)

Análise do transcrito do mexilhão marrom (*Perna perna*) sob
contaminação por antraceno

Jhonatas Sirino Monteiro

São Paulo
2017

Jhonatas Sirino Monteiro

**Análise do transcrito do mexilhão marrom (*Perna perna*) sob
contaminação por antraceno**

Dissertação apresentada ao Instituto de
Matemática e Estatística da Universidade
de São Paulo como requisito para obtenção
do título de Mestre em Ciências.

Área de concentração: Bioinformá-
tica

Orientador: Dr. João Carlos Setubal

São Paulo
2017

Monteiro, Jhonatas S.

Análise do transcrito do mexilhão marrom (*Perna perna*)
sob contaminação por antraceno

87 páginas

Dissertação (Mestrado) - Instituto de Matemática e Estatística da Universidade de São Paulo.

1. Bioinformática
2. Transcritômica
3. Ecotoxicologia

I. Universidade de São Paulo. Instituto de Matemática e Estatística.

*Dedico à minha esposa, Vanessa, à minha família,
aos amigos e aos professores que tornaram esse
sonho possível!*

AGRADECIMENTOS

Gostaria de agradecer primeiramente minha esposa, Vanessa, que me apoiou em todos os aspectos ao longo de todo o meu trajeto, permitindo que me dedicasse exclusivamente a essa pesquisa. Além de aturar os meus momentos de angústia, nervosismo e chatices. Você é o pilar que sustenta todo esse trabalho e toda a minha vida.

Quero agradecer igualmente a minha família que me ensinaram os verdadeiros valores da vida, além de me apoiarem todas as minhas escolhas, o que me permitiu estar aqui hoje apresentando este trabalho. Estar longe de vocês é difícil, mas o amor e carinho de vocês, mesmo a distância, foram imprescindíveis para o desenvolvimento deste trabalho.

Quero agradecer aos amigos e colegas de equipe que deram suporte não só em questões técnicas e científicas, mas também no apoio emocional. Obrigado por compartilharem suas experiências de vida comigo.

Agradeço a CAPES pela bolsa de mestrado (projeto BIGA, edital 51/2013, Programa de Biologia Computacional) e aos colaboradores que permitiram a execução de cada etapa deste trabalho.

Agradeço também ao Prof. Dr. João Carlos Setubal que abriu as portas de seu laboratório e me deu a chance de aprender o que é ser bioinformata. Seus ensinamentos serão sempre lembrados e espero que com esse trabalho eu possa retribuir, pelo menos um pouco, a oportunidade que você me deu de estar aqui e a importância que isso representa para a minha vida. Serei eternamente grato à você Professor.

E por fim, um agradecimento especial ao Prof. Dr. Igor Dias Medeiros, que não só foi mentor da presente pesquisa que hoje apresento, mas foi um grande amigo. Se hoje estou trilhando o caminho da ciência foi porque você me fez acreditar que eu podia seguir esse caminho. Agradeço toda a sua dedicação em me ensinar o que é fazer ciência, e também todo o seu apoio que tive tanto na minha vida pessoal e profissional. Espero que esse trabalho também retribua, pelo menos um pouco, tudo que você fez por mim.

RESUMO

O mexilhão marrom *Perna perna* (Linnaeus, 1758) auxilia no monitoramento de compostos químicos em ecossistemas marinhos. No entanto, os mecanismos moleculares de detoxificação e resposta ao estresse são desconhecidos. Elucidar esses mecanismos é crucial para entender os efeitos tóxicos dos poluentes químicos e desenvolver biomarcadores para avaliar a qualidade ambiental dos ecossistemas marinhos. No presente estudo, indivíduos da espécie *P. perna* foram expostos a antraceno (ANT) e os RNAs mensageiros (mRNA) das brânquias foram sequenciados com a plataforma Illumina. A análise química do tecido mole dos animais identificou concentrações de ANT 268 a 715 vezes mais alta no grupo exposto comparado ao grupo controle, demonstrando que a exposição foi realizada com sucesso.

O sequenciamento do transcrito do *P. perna* gerou 273.152.390 pares de *reads*, resultando na montagem de 231.728 *contigs* com tamanho médio de 720 pb e N50 de 1.083 pb, os quais 66.563 *contigs* (28,7%) pode ser anotado utilizando banco de dados como GenBank, Pfam, Gene Ontology e KEGG. Os resultados obtidos a partir da anotação funcional sugerem que as brânquias tenham papel na biotransformação de xenobióticos, resposta antioxidante, sinalização, resposta imunológica inata, e osmorregulação.

Foi possível identificar genes de biotransformação de fase I, II e III, incluindo CYPs e GSTs. Transcritos similares a CYPs e GSTs estavam sendo expressos no grupo exposto, porém nenhum deles foram classificados como diferencialmente expressos. Contudo, muitos genes hipotéticos foram diferencialmente expressos, o que sugere que *P. perna* utilize mecanismos desconhecidos de biotransformação para lidar com a contaminação de ANT. Genes de sistema imune inato foram regulados tanto positivamente quanto negativamente, assim como observado para *Perna viridis* exposto a benzo(a)pireno, sugerindo que ANT promove alterações da capacidade de resposta do sistema imune inato do *P. perna*.

ABSTRACT

The brown mussel *Perna perna* (Linnaeus, 1758) helps the monitoring of chemical compounds in marine ecosystems. However its molecular mechanisms of detoxification and stress response remain unclear. Elucidating these mechanisms is crucial to understand the toxic effects of chemical pollutants and to develop biomarkers to assess marine ecosystems. In this study, *P. perna* individuals were exposed to anthracene (ANT) and its mRNA complement was sampled sequenced with Illumina technology. Chemical analysis of the soft tissue identified ANT concentrations 268 - 715 fold higher in the exposed group compared to controls, demonstrating that the exposure procedure was successfully accomplished.

Transcriptome sequencing of *P. perna* generated 273.152.390 paired reads that were assembled in 231.728 contigs of average length 720 bp and N50 1.083 bp , which 66.563 contigs (28,7%) could be annotated using GenBank genes, Pfam domains, Gene Ontology (GO) terms and KEGG pathways. The results obtained from functional annotation suggest gills play a role in xenobiotics biotransformation, antioxidant response, signal transduction, innate immune response, and osmoregulation.

It was possible to identify transcripts similar to genes related with biotransformation reactions of phases I, II and III, including CYPs and GSTs. Transcripts similar to CYPs and GSTs isoforms were highly expressed in the group exposed to ANT, however no CYP, GST, or even other genes related with biotransformation reactions were classified as differentially expressed. On the other hand, several hypothetical genes were differentially expressed, which suggests that *P. perna* uses unknown mechanisms of biotransformation to deal with ANT stress contamination. Immune related-genes were both up and down-regulated, as was also observed for *Perna viridis* exposed to benzo(a)pyrene, suggesting that ANT promotes alteration in the immune response of *P. perna*.

Lista de Figuras

1	Quantidade de dados registrados no NCBI para as espécies: a) <i>Perna perna</i> ; b) <i>Mytilus galloprovincialis</i> ; C) <i>Mytilus edulis</i> d) <i>Crassostrea gigas</i> . Dados colhidos no dia 23, Nov de 2017.	22
1.1	<i>Pipeline</i> de bioinformática genérico de análise de expressão diferencial com dados de RNA-Seq.	27
2.1	Reconstrução da filogenia do <i>P. perna</i> utilizando o gene 18S rRNA. Os números de suporte dos nós representam os valores de <i>bootstrap</i> , e a escala representa o número de substituições de nucleotídeo por sítio. As cores dos clados representam as famílias: vermelho = <i>Mytilidae</i> , azul = <i>Pteriidae</i> , Cinza = <i>Ostreidae</i> . O clado em preto representa o grupo externo	35
2.2	Distribuição dos tamanhos dos <i>contigs</i> montados. O gráfico a) apresenta todos os tamanhos, enquanto o gráfico b) é um recorte do gráfico a) limitando o tamanho máximo em 6.000 pb.	36
2.3	Curva de rarefação feita a partir do alinhamento das <i>reads</i> com o transcrito montado	38
2.4	Os 20 termos GO mais abundantes na categoria BP a nível de aminoácido. O eixo X representa o número de transcritos associados	40
2.5	Os 20 termos GO mais abundantes na categoria MF a nível de aminoácido. O eixo X representa o número de transcritos associados	41
2.6	Anotação das proteínas, preditas a partir das sequências de transcritos, com o banco de vias do KEGG.	42
2.7	As vias de sistema imune anotadas com o banco de dados KEGG.	43
2.8	Genes do sistema imune inato encontrados em <i>P. perna</i> comparando com alguns organismos modelos. O número de genes das espécies modelo foram retirados do artigo do Guo et al [59]. Os números para <i>P. perna</i> representam os <i>unigenes</i> montados pelo Trinity, e portanto, esses valores são aproximações do número real de genes que se espera encontrar.	44
3.1	Comparações dos números absolutos de transcritos similares a membros da família CYP identificados em <i>P. perna</i> com outros bivalves.	46
3.2	Comparações dos números de transcritos similares a membros da família CYP identificados em <i>P. perna</i> com outros bivalves, normalizados pelo total de transcritos identificados na espécie.	47
3.3	Níveis de expressão, quantificados em TPM pelo Kallisto, das 10 isoformas de CYPs mais expressas identificadas em <i>P. perna</i> em ambos os grupos. . .	48
3.4	Comparações dos números absolutos de GSTs identificadas em <i>P. perna</i> com outros bivalves.	48

3.5	Comparações dos números absolutos de GSTs identificadas em <i>P. perna</i> com outros bivalves, normalizados pelo total de transcritos identificados na espécie.	49
3.6	Níveis de expressão, quantificadas em TPM pelo Kallisto, das 10 isoformas de GSTs mais expressas identificadas em <i>P. perna</i> em ambos os grupos. . .	49
3.7	Relação da média com a variância obtida da estimação dos níveis de expressão dos transcritos. A linha azul representa o modelo de distribuição binomial negativa, a linha preta o modelo de distribuição de Poisson, e os pontos cinzas representam os dados observados.	50
3.8	Diagrama de Venn dos TDEs identificados a partir dos dados de estimativas do Salmon e Kallisto.	51
3.9	<i>Volcano plot</i> dos TDEs identificados a partir dos dados de estimativas do Salmon e Kallisto.	51
3.10	MDS plot das réplicas utilizadas no estudo. A1, A2, A3 e A4 representam as réplicas do grupo exposto a ANT, enquanto C1, C2, C3 e C4 representam as réplicas do grupo controle.	54
3.11	PCA plot das réplicas do grupo exposto.	55
3.12	PCA plot das réplicas do grupo controle.	55
A1	Análise da qualidade do sequenciamento por base da <i>read</i> 1, realizado pelo FastQC da amostra 1 do grupo exposto.	68
A2	Análise da qualidade do sequenciamento por base da <i>read</i> 2, realizado pelo FastQC da amostra 1 do grupo exposto.	69
A3	Análise da qualidade do sequenciamento por base da <i>read</i> 1, realizado pelo FastQC da amostra 2 do grupo exposto.	70
A4	Análise da qualidade do sequenciamento por base da <i>read</i> 2, realizado pelo FastQC da amostra 2 do grupo exposto.	71
A5	Análise da qualidade do sequenciamento por base da <i>read</i> 1, realizado pelo FastQC da amostra 3 do grupo exposto.	72
A6	Análise da qualidade do sequenciamento por base da <i>read</i> 2, realizado pelo FastQC da amostra 3 do grupo exposto.	73
A7	Análise da qualidade do sequenciamento por base da <i>read</i> 1, realizado pelo FastQC da amostra 4 do grupo exposto.	74
A8	Análise da qualidade do sequenciamento por base da <i>read</i> 2, realizado pelo FastQC da amostra 4 do grupo exposto.	75
A9	Análise da qualidade do sequenciamento por base da <i>read</i> 1, realizado pelo FastQC da amostra 1 do grupo controle.	76
A10	Análise da qualidade do sequenciamento por base da <i>read</i> 2, realizado pelo FastQC da amostra 1 do grupo controle.	77
A11	Análise da qualidade do sequenciamento por base da <i>read</i> 1, realizado pelo FastQC da amostra 2 do grupo controle.	78
A12	Análise da qualidade do sequenciamento por base da <i>read</i> 2, realizado pelo FastQC da amostra 2 do grupo controle.	79
A13	Análise da qualidade do sequenciamento por base da <i>read</i> 1, realizado pelo FastQC da amostra 3 do grupo controle.	80
A14	Análise da qualidade do sequenciamento por base da <i>read</i> 2, realizado pelo FastQC da amostra 3 do grupo controle.	81
A15	Análise da qualidade do sequenciamento por base da <i>read</i> 1, realizado pelo FastQC da amostra 4 do grupo controle.	82

A16	Análise da qualidade do sequenciamento por base da <i>read</i> 2, realizado pelo FastQC da amostra 4 do grupo controle.	83
B1	Os 20 termos GO mais abundantes na categoria BP a nível de nucleotídeo. O eixo X representa o número de transcritos associados	85
B2	Os 20 termos GO mais abundantes na categoria MF a nível de nucleotídeo. O eixo X representa o número de transcritos associados	86
B3	Os 20 termos GO mais abundantes na categoria CC a nível de nucleotídeo. O eixo X representa o número de transcritos associados	86
B4	Os 20 termos GO mais abundantes na categoria CC a nível de proteína. O eixo X representa o número de transcritos associados	87

Lista de Tabelas

2.1	Números de acesso das sequências do gene 18S rRNA depositadas no Gen-Bank utilizadas para reconstrução da filogenia	33
2.2	Concentração de ANT (ng/g) no tecido mole do <i>P. perna</i> após 24h de exposição.	35
2.3	Estatísticas gerais sobre a montagem <i>de novo</i> do transcrito do <i>P. perna</i> . As amostras C1, C2, C3 e C4 representam as réplicas do grupo controle. enquanto as amostras A1, A2, A3 e A4 representam as réplicas do grupo exposto.	37
2.4	Relatório da análise de integridade da montagem do transcrito do <i>P. perna</i> realizada com o BUSCO.	38
2.5	Anotação funcional dos <i>contigs</i> . Pfam (nucleotídeo) são resultados obtidos a partir do Interproscan onde o arquivo de entrada continha sequências de nucleotídeos. Pfam (aminoácido) são resultados obtidos a partir do HMMER onde o arquivo de entrada continha sequências de aminoácidos .	39
3.1	Os 20 TDEs mais expressos/inibidos em <i>P. perna</i> após 24h de exposição ao ANT.	52

Sumário

Introdução	21
1 Objetivos, definições e trabalhos relacionados	25
1.1 Objetivo do trabalho	25
1.1.1 Objetivos específicos	25
1.2 Conceitos e definições	25
1.2.1 Biomarcadores	25
1.2.2 Biotransformação	26
1.2.3 <i>Pipeline</i> de bioinformática para montagem e análise do transcrito	26
1.3 Trabalhos relacionados	28
1.3.1 <i>De novo</i> transcriptome analysis of <i>Perna viridis</i> highlights tissue-specific patterns for environmental studies	29
1.3.2 Transcriptomic responses of <i>Perna viridis</i> embryo to benzo(a)pyrene exposure elucidated by RNA sequencing	30
2 Montagem e anotação funcional do transcrito	31
2.1 Metodologia	31
2.1.1 Exposição dos mexilhões	31
2.1.2 Extração do RNA e sequenciamento	31
2.1.3 Análise química do tecido	32
2.1.4 Reconstrução da filogenia do mexilhão	32
2.1.5 Montagem <i>de novo</i> do transcrito	32
2.1.6 Descontaminação dos dados	34
2.1.7 Anotação funcional	34
2.2 Resultados e Discussões	34
2.2.1 Filogenia do <i>P. perna</i>	34
2.2.2 Concentração de ANT no tecido mole	35
2.2.3 Sequenciamento e montagem do transcrito	36
2.2.4 Anotação funcional do transcrito	39
3 Genes de biotransformação de HPAs e análise de expressão diferencial	45
3.1 Metodologia	45
3.1.1 Estimativa dos níveis de expressão gênica	45
3.1.2 Cálculo da expressão diferencial	45
3.2 Resultados e Discussões	46
3.2.1 Transcritos relacionados com biotransformação de HPA	46
3.2.2 Análise de expressão diferencial	50
Conclusão	57
Bibliografia	58
Anexo A	67
Anexo B	85

Introdução

A contaminação de ecossistemas por hidrocarbonetos do petróleo, presentes nos combustíveis fósseis e nos lubrificantes, é um risco presente no cenário mundial devido à forte presença de atividades humanas que fazem uso desses compostos.

Dentre os hidrocarbonetos do petróleo, existem os hidrocarbonetos policíclicos aromáticos (HPAs), que são contaminantes comuns em ambientes aquáticos [1], com uma preocupação especial para ecossistemas costeiros e marinhos devido a suscetibilidade à contaminação provocada por atividades portuárias, refinarias costeiras e áreas industriais associadas a grandes cidades em zonas costeiras [2], como por exemplo a região da Baixada Santista e o litoral norte do estado de São Paulo.

A Baixada Santista abriga o Sistema Estuarino Santos-São Vicente que por sua vez comporta o maior porto da América Latina, o porto de Santos, e o polo industrial da cidade de Cubatão, enquanto no litoral norte se encontra o Terminal Aquaviário de São Sebastião (Tebar), "o maior terminal da América do Sul, responsável por 50% do volume de petróleo processado no país" [3], sendo essas possíveis fontes de contaminação por HPAs nos ecossistemas marinhos locais. Nessas regiões se desenvolvem também atividades que são conflitantes com esse cenário, como o turismo e a pesca. Um estudo realizado por [4] confirmou a presença de HPAs nos sedimentos e em ostras (*Crassostrea rhizophorae*) no estuário de Santos, e dentre os HPAs encontrados foi identificado a presença do antraceno (ANT). O ANT é um composto de três anéis benzênicos de baixo peso molecular, e se encontra na lista de prioridade química da *United States Environmental Protection Agency* (EPA) [5].

A preocupação envolvida na presença de ANT no ambiente decorre de sua toxicidade. Vieira et. al. [6] relatam os efeitos do ANT em *Pomatoschistus microps*, um peixe estuarino. Nesta espécie, o contaminante provocou alterações como inibição da acetilcolinesterase, e, ainda, o induziu a sofrer estresse oxidativo. Adicionalmente, o baixo peso molecular do ANT confere maior solubilidade em água, fazendo com que ele se encontre em maiores concentrações na coluna d'água [7], e portanto mais biodisponível, e ainda sua toxicidade pode ser aumentada pela exposição à luz solar, o que é abundante em regiões tropicais [8, 9]. Assim, é imprescindível o monitoramento da presença de contaminantes para auxiliar na tomada de medidas mitigadoras.

Os bivalves desempenham um papel importante em estudos de qualidade ambiental, pois apresentam uma distribuição global, habilidade de bioacumular contaminantes [1], e também por serem sésseis, sendo, portanto, representantes fiéis das condições ambientais locais. O mexilhão marrom (*Perna perna*) é um dos bivalves utilizados para estudos de monitoramento da qualidade ambiental [10].

O *P. perna* pode ser encontrado na costa do continente africano, e também ao longo da costa Atlântica do continente sul-americano, em maior abundância no Brasil entre os estados do Espírito Santo e Santa Catarina [11, 12]. Sabe-se que o *P. perna* tem um importante papel na estruturação da comunidade e diversidade dos ecossistemas marinhos,

além de sua importância econômica, já que esse animal é comercializado para consumo humano, e também devido a seu rápido crescimento e alta resistência a variações físico-químicas da água, características ideais para a aquicultura [12].

Uma das formas de se monitorar a qualidade dos ecossistemas marinhos é através do uso de biomarcadores. Em estudos com o *P. perna*, têm-se medido a atividade de enzimas de biotransformação de xenobióticos (glutathione-S-transferase (GST) e proteína de resistência a múltiplas drogas (MDR)) e antioxidantes (glutathione peroxidase (GPx), catalase (CAT) e superóxido dismutase (SOD)), como biomarcadores [10, 13, 14]. No entanto, a atividade enzimática apresenta respostas com um atraso maior, em relação ao momento que houve a exposição ao contaminante, do que a mudança do padrão de expressão gênica, como mostra um estudo de Lüchmann et al. [1], onde ostras (*Crassostrea brasiliana*) foram expostas a fenantreno por 24h e apresentaram alterações no padrão de expressão de genes relacionados a biotransformação de HPAs, mas não houve uma alteração significativa na atividade dessas enzimas, o que nos mostra que alterações na expressão gênica são biomarcadores de resposta mais rápida do que atividade enzimática.

Dentre os tecidos dos bivalves, estudos tem apresentado evidências que as brânquias possuem um papel no processo de biotransformação de alguns xenobióticos[1, 15]. Este tecido é um dos primeiros tecidos a ter contato com o ambiente contaminado, responsável por filtrar a água, e possivelmente o primeiro a sofrer alterações na expressão de genes. Portanto, as brânquias podem ser uma boa opção para uma primeira investigação dos mecanismos moleculares que são utilizados pelo *P. perna* para lidar com ANT.

Do ponto de vista da genômica e transcritômica, pouco se conhece sobre o *P. perna* se comparado a outros bivalves, como a ostra do pacífico (*Crassostrea gigas*) e os mexilhões *Mytilus galloprovincialis* e *Mytilus edulis* (Figura 1).

a)		b)		c)		d)	
Entrez records		Entrez records		Entrez records		Entrez records	
Database name	Direct links	Database name	Direct links	Database name	Subtree links	Database name	Direct links
Nucleotide	1,487	Nucleotide	161,157	Nucleotide	3,925	Nucleotide	71,542
Protein	1,206	Nucleotide EST	19,617	Nucleotide EST	5,315	Nucleotide EST	206,388
Genome	1	Nucleotide GSS	1	Protein	1,997	Protein	75,094
Popset	20	Protein	5,936	Structure	9	Structure	5
PubMed Central	68	Structure	6	Genome	1	Genome	1
Gene	12	Genome	1	Popset	79	Popset	63
Taxonomy	1	Popset	78	Conserved Domains	1	GEO Datasets	807
		GEO Datasets	419	GEO Datasets	81	PubMed Central	1,284
		PubMed Central	695	PubMed Central	1,040	Gene	36,468
		Gene	12	Gene	12	SRA Experiments	1,780
		SRA Experiments	103	SRA Experiments	188	Protein Clusters	12
		Protein Clusters	12	Protein Clusters	12	Bio Project	75
		Bio Project	35	Bio Project	15	Bio Sample	1,877
		Bio Sample	121	Bio Sample	191	Bio Systems	261
		Assembly	2	Probe	4	Assembly	1
		Probe	20	PubChem BioAssay	6	Probe	115
		Taxonomy	1	Taxonomy	2	PubChem BioAssay	1
						Taxonomy	1

Figura 1: Quantidade de dados registrados no NCBI para as espécies: a) *Perna perna*; b) *Mytilus galloprovincialis*; c) *Mytilus edulis*; d) *Crassostrea gigas*. Dados colhidos no dia 23, Nov de 2017.

As tecnologias de sequenciamento de nova geração, conhecidas em inglês como *next-generation sequencing* (NGS), permitiram gerar transcritomas de organismos não-modelos com menor custo e de forma mais acessível. O RNA-Seq é um dos métodos de sequenciamento em larga escala que tem sido utilizado para análise de expressão gênica em organismos não-modelos [16]. A análise do transcrito do *P. perna* permitirá não apenas observar os principais mecanismos de biotransformação, podendo determinar quais genes são potenciais biomarcadores de exposição sensíveis à presença de ANT no ambiente, mas irá permitir conhecer também outras funções biológicas da espécie que continuam desconhecidas, pois, para o conhecimento do autor, este é o primeiro estudo com sequenciamento de RNA em larga escala feito para esse mexilhão em condições laboratoriais. Atualmente existe um estudo em andamento, com *Perna perna* selvagens, em parceria UNESP-UNIFESP-UFSC e Universidade de Padova, sendo um dos responsáveis pelo estudo o Prof. Dr. Denis Moledo de Souza Abessa, da UNESP.

Ao longo desse documento serão apresentados os objetivos traçados para execução do presente trabalho; os conceitos de biomarcador e biotransformação, cujas definições são importantes para contextualização da pesquisa; explanação das etapas seguidas em um pipeline de análise de transcrito de uma espécie não-modelo; e um breve resumo de trabalhos relacionados publicados recentemente. Todos os itens anteriores estão compreendidos no primeiro capítulo. No capítulo seguinte serão apresentados os métodos utilizados para realizar o experimento de exposição ao ANT, a montagem e a anotação funcional do transcrito, assim como os resultados discutidos obtidos pela metodologia aplicada. E no último capítulo serão apresentados os transcritos que possivelmente estão relacionados com a biotransformação de ANT, e também a metodologia aplicada para a comparação entre os grupos definidos no desenho experimental, assim como os resultados obtidos dessa comparação.

Este projeto foi formulado principalmente com base nas ideias do prof. Dr. Igor Dias Medeiros, da UNIFESP-Baixada Santista, que foi orientador de iniciação científica do presente autor. O prof. Igor foi co-orientador informal do presente trabalho.

Capítulo 1

Objetivos, definições e trabalhos relacionados

1.1 Objetivo do trabalho

Montagem e análise do transcrito de *Perna perna* exposto a ANT e identificação dos genes induzidos ou inibidos durante a exposição.

1.1.1 Objetivos específicos

1. Definir o delineamento experimental geral para alcançar o objetivo geral proposto;
2. Sequenciar mRNA de *Perna perna* de acordo com o delineamento experimental definido em (1);
3. Definir e implementar um *pipeline* de análises da bioinformática;
4. Utilizar *pipeline* definido em (3) para montar o transcrito utilizando os dados gerados em (2);
5. Determinar os transcritos diferencialmente expressos (TDEs) em função da exposição ao ANT, realizar sua classificação funcional e interpretar os resultados.

1.2 Conceitos e definições

1.2.1 Biomarcadores

Desde o surgimento do termo em 1987, a definição de biomarcador evoluiu até ser hoje definido como “respostas biológicas adaptativas a estressores, evidenciadas como alterações bioquímicas, celulares, histológicas, fisiológicas ou comportamentais” [17].

Segundo Zagatto e Bertoletti [17], os biomarcadores são classificados em 3 tipos:

- **Exposição:**

Indicam a biodisponibilidade de contaminantes ou de seus metabólitos e se correlacionam com sua concentração no meio ambiente;

- **Efeito:**

Indicam a ocorrência de um estresse, caracterizado pela indução de mecanismos de defesa a nível molecular-bioquímico, que poderá ser reversível assim que estressor parar de agir sobre o organismo;

- **Suscetibilidade:**

Indicadores de processos que causam variações de respostas ao longo do tempo e entre exposição e efeito, determinando condições que aumentam a taxa de transição entre esses dois extremos.

Alguns biomarcadores, acetilcolinesterase (AChE) por exemplo, se encaixam tanto na categoria de exposição quanto na de efeitos, e por isso, alguns autores defendem uma quarta categoria, que seria biomarcadores de exposição e efeito [17]. Nem toda resposta biológica é considerada um biomarcador. Para tal, faz-se necessário seguir alguns critérios: "facilidade de medida, resposta-dose ou tempo-dependente ao contaminante, compreensão e especificação de limites de variabilidade devida a outros fatores que não a ação do contaminante, e significância biológica"[17].

1.2.2 Biotransformação

Biotransformação é processo de alterações químicas estruturais, mediado por enzimas, que tem como principal função transformar substâncias lipofílicas em derivados hidrofílicos e com menor toxicidade, apesar do último nem sempre ocorrer [17, 18].

A biotransformação é dividida em três fases, mas apenas didaticamente, pois algumas substâncias passam por apenas uma das fases [17]:

- **Fase I:**

Caracterizada pela adição de um grupamento funcional que facilita as reações da fase II [17]. Geralmente, o grupo adicionado confere maior polaridade à substância, tornando-a mais hidrossolúvel;

- **Fase II:**

Conjugação do xenobiótico com algum componente do tecido normal ou substrato endógeno, resultando em um complexo de menor toxicidade e ionizável, que visa facilitar sua excreção [17, 18];

- **Fase III:**

Envolve transportadores de membrana, que tem como principal função levar os conjugados da fase II e metabólitos para o meio extracelular, mas também possuem papéis importantes na absorção e distribuição de alguns xenobióticos, ou ainda atuar como uma barreira contra a entrada de xenobióticos na célula [19].

1.2.3 Pipeline de bioinformática para montagem e análise do transcrito

Estudos no âmbito da transcriptômica, em geral, tendem a seguir etapas semelhantes até ser feita a reconstrução do transcrito. Os passos posteriores são determinados pela pergunta/hipótese que motivou o estudo. Abaixo se encontra um *pipeline* de bioinformática que visa a análise de expressão gênica diferencial (Figura 1.1).

Figura 1.1: *Pipeline* de bioinformática genérica de análise de expressão diferencial com dados de RNA-Seq.

1. Análise da qualidade do sequenciamento

Caracteriza-se por verificar os *phred scores* [20] obtidos em cada posição das sequências, número de *reads* sequenciadas, conteúdo GC, identificar sequências contaminantes, adaptadores da plataforma de sequenciamento, cauda poli A/T e sequências super-representadas, além de outras estatísticas. Essa análise permite ajustar os parâmetros dos softwares da etapa seguinte, permitindo extrair o máximo de dados para a montagem do transcriptoma.

2. Filtragem

Ocorre a remoção, completa ou parcial, de *reads* de baixa qualidade, contaminantes, adaptadores, cauda poli A/T, entre outros parâmetros que podem comprometer a etapa seguinte, a montagem.

3. Montagem do transcriptoma

Ocorre a reconstrução dos transcritos a partir das *reads* sequenciadas. Entre as abordagens existentes, as duas mais utilizadas são: montagem referenciada e a montagem *de novo*. A primeira é utilizada quando o organismo estudado possui o seu genoma (preferencial) ou transcriptoma sequenciado. Nesse caso, os softwares de montagem primeiro alinham as *reads* com a referência para se obter as coordenadas do alinhamento, e, então, as *reads* que alinham em um mesmo *locus* gênico são utilizadas para a construção do *contig*. Já a segunda é recomendada para organismos que não possuem um genoma ou transcriptoma sequenciado, ou para genomas muito fragmentados. Os softwares existentes utilizam diferentes estratégias para a reconstrução dos transcritos, mas a mais frequente é utilizando estrutura de grafos [21, 22, 23, 24].

4. Anotação dos genes/transcritos

Ocorre a tentativa de identificação e classificação funcional dos genes/transcritos montados. Nessa etapa são feitas buscas por similaridade com outras sequências já conhecidas depositadas em bancos de dados públicos, buscas por domínios conservados de proteínas, classificação funcional utilizando os termos do Gene Ontology (GO), identificação das vias metabólicas que os genes/transcritos participam, entre outras técnicas de anotação que visam caracterizar os genes/transcritos encontrados no estudo.

5. Quantificação da expressão gênica

É medido quanto um determinado gene/transcrito foi expresso pelo indivíduo em um determinado contexto. Essa medida é obtida a partir da contagem do n° de *reads* sequenciadas pertencentes a um gene/transcrito. Para isso, o primeiro desafio é determinar a qual transcrito cada *read* pertence.

As primeiras ferramentas faziam o alinhamento das *reads* com o transcrito recém-montado, associando as *reads* com seus respectivos transcritos, e então era feita a contagem. O problema dessa primeira abordagem ocorre quando uma mesma *read* é alinhada com mais de um transcrito (*multi-mapping*), o que gera uma dificuldade ao quantificar a abundância e introduz alguns vieses nos resultados, além do grande tempo e recursos consumidos para processar todos os alinhamentos [25]. Para superar tais barreiras alguns métodos descartavam todos os *multi-mapping reads* [26], mas havia perda de dados que poderiam ser importantes para análises posteriores. Então se passou a utilizar algumas estratégias que levavam em conta os *multi-mapping reads*, estimando a abundância relativa usando a função de máxima verossimilhança com o algoritmo maximização de expectativas (EM) e variações desse algoritmo para otimizar a iteração da função, mas ainda havia o limitante do tempo e recurso gastos com o alinhamento das *reads* [25].

Os mais recentes softwares trouxeram uma nova abordagem chamada “*alignment free*”, em que as ferramentas não mais realizam um alinhamento completo das *reads* com a referência, mas utilizam outros métodos como “*pseudoalignment*” do Kallisto [27], *lightweight-alignment* do Salmon [28] e *quasi-mapping* do RapMap [29], hoje também implementado pelo Salmon. Essas novas abordagens reduziram o tempo e recursos gastos nas análises, além de aumentar a acurácia das estimativas de abundância [25, 29].

6. Análise de expressão gênica diferencial

Ocorre a comparação entre 2 ou mais grupos amostrados a fim de se identificar quais os genes/transcritos foram expressos com uma diferença significativa entre os grupos. Atualmente existem diversos pacotes estatísticos que implementam diferentes métodos para identificar genes/transcritos diferencialmente expressos, e cada um possui vantagens e desvantagens frente aos diversos *datasets* de RNA-Seq, sendo difícil prever qual deles é o mais indicado. Seyednasrollah et al. [30] apresenta uma comparação das ferramentas de análise de expressão diferencial que representavam o estado da arte até o ano de publicação do artigo. Esses estudos ajudam na escolha dos *softwares* mais adequados ao objeto de estudo.

1.3 Trabalhos relacionados

Lüchmann et al. [1] expôs ostras da espécie *Crassostrea brasiliana* a diferentes concentrações de fenantreno e sequenciou o transcrito, identificando 6 novos possíveis genes de CYP (citocromo P450) e 4 GSTs (glutathione S-transferase) para a espécie.

Wong et al. [31] realizaram a análise do transcrito do mexilhão de mar profundo, *Bathymodiolus platifrons*, que habita fontes hidrotermais, “cold seeps” e outros habitats ricos em matéria orgânica, buscando por genes relacionados a resposta do sistema imune e mecanismos de detoxificação.

Moreira et al. [32] estudaram o transcrito do mexilhão *Mytilus galloprovincialis*

comparando os perfis de expressão gênica entre os diferentes tecidos do animal, identificando os genes responsáveis nos mecanismos de resposta do sistema imune e fisiológico.

Xu et al. [33] analisaram o transcrito do mexilhão *Mytilus coruscus*, buscando por genes que respondam ao estresse por contaminação por metal pesado, o cobre. O objetivo foi descobrir quais os mecanismos utilizados pelo animal para lidar com o estresse a nível molecular.

Como é possível notar, existe interesse global em conhecer o transcrito de bivalves marinhos, pois como os próprios estudos justificam, são animais modelo para diversos estudos de impactos ambientais e muitas espécies possuem valor comercial, além de terem também importância ecológica, visto que são consumidores primários e servem de alimento para predadores. Conhecer seus mecanismos de defesa à nível molecular permite desenvolver inúmeras aplicações que podem auxiliar no monitoramento ambiental, ou ainda, no cultivo da espécie na aquicultura.

A seguir são apresentados, resumidamente, dois estudos de análise de transcrito do mexilhão *Perna viridis* expostos a contaminantes ambientais, incluindo HPAs, e outros estressores.

1.3.1 *De novo* transcriptome analysis of *Perna viridis* highlights tissue-specific patterns for environmental studies

Leung et. al. [15] sequenciaram e analisaram o transcrito do mexilhão verde, *Perna viridis*, espécie utilizada como biomonitor da região Indo-Pacífico para monitoramento ambiental e investigações ecotoxicológicas.

Mexilhões adultos foram expostos a múltiplos estressores (metais, desreguladores endócrinos, poluentes orgânicos e nano partículas) e a diferentes condições do meio físico (temperatura, salinidade, níveis de oxigênio dissolvido e pH) durante 24h. Os animais então foram dissecados e amostras dos tecidos do músculo adutor (ma), brânquias (bq) e hepatopâncreas (hp) foram colhidas. O RNA total então foi extraído de 6 amostras de tecido (macho-ma, fêmea-ma, macho-bq, fêmea-bq, macho-hp e fêmea-hp), sendo construída uma biblioteca de cDNA, a partir do mRNA purificado do RNA total, para cada amostra. As bibliotecas foram sequenciadas individualmente em 6 *lanes* da plataforma Illumina Genome Analyzer IIX.

Ao todo foram sequenciados aproximadamente 544 milhões de *reads*, submetidos a montagem *de novo* com o *software* Trinity. Os *contigs* gerados pelo Trinity que tinham 95% de sua sequência compartilhada com outros *contigs* foram agrupadas para diminuir redundância, gerando um total de 192.879 de *contigs*, com tamanho médio de 1.264 pb, e N50 de 2.868 pb. Os *contigs* então foram submetidos a processos de anotação funcional, utilizando buscas de similaridade com bancos de dados do NCBI, busca por domínio de famílias de proteínas e termos do Gene Ontology (GO) associados a esses domínios, e buscas por vias metabólicas envolvidas com o banco de dados do KEGG. A expressão gênica também foi quantificada.

Os termos do GO mais abundantes foram "*metabolic process*" na categoria "*biological process*", "*binding*" em "*molecular function*" e "*cell*" em "*cellular component*". Análises multivariadas mostraram que os perfis de expressão gênica não variaram entre os sexos, mas sim entre os tipos de tecidos amostrados. Em contrapartida, a composição de termos do GO não variou em nenhuma das comparações.

Entre os *contigs* anotados, 187 foram associados ao termo "*response to stress*" e 96 estavam relacionados com a via "*xenobiotics biodegradation and metabolism*". *Contigs*

relacionados com resposta ao estresse foram mais expressos no hepatopâncreas, enquanto os relacionados com biodegradação e metabolismo de xenobióticos foram mais expressos nas brânquias e hepatopâncreas. Entre os *contigs* relacionados com resposta ao estresse, a família de proteínas de choque térmico (HSPs) foi uma das mais abundantes, enquanto as famílias citocromo P450 (CYPs) e glutathione S-transferase (GSTs) foram as mais abundantes entre os *contigs* da via de biodegradação e metabolismo de xenobióticos.

1.3.2 Transcriptomic responses of *Perna viridis* embryo to benzo(a)pyrene exposure elucidated by RNA sequencing

Jiang et. al. [34] expuseram embriões de mexilhões *Perna viridis* a benzo(a)pireno diluído em sulfóxido de dimetilo (DMSO) por 24h. O RNA total foi extraído, isolando o mRNA para construção das bibliotecas de cDNA e sequenciamento pela plataforma Illumina HiSeqTM2000. Ao todo foram sequenciados, aproximadamente, 93 milhões de *reads*, gerando 99.202 *contigs* no processo de montagem com o Trinity, com tamanho médio de 981 pb e N50 de 2383 pb.

Os *contigs* foram anotados utilizando bancos de dados do NCBI, SwissProt, COG. Os agrupamentos funcionais mais presentes no banco de dados do COG foram "*general function prediction*", "*translation*", "*ribosomal structure*" e "*biogenesis*". Apenas os genes diferencialmente expressos (GDEs) foram anotados utilizando os bancos de dados do GO e KEGG. Os termos mais presentes foram: "*metabolic process*" e "*cellular process*" na categoria "*biological process*", "*cell*", "*cell part*" e "*organelle*" na categoria "*cellular component*", e "*binding*" e "*catalysis*" na categoria "*molecular function*".

A análise de enriquecimento de vias do KEGG mostrou que os genes que foram mais expressos estavam relacionados com vias do sistema imune "*antigen processing and presentation*", "*phagosome*" e "*protein processing in endoplasmic reticulum*", enquanto os genes mais inibidos estavam relacionados com as vias "*ECM-receptor interaction*"; e "*Complement and coagulation cascades*". Entre os genes mais expressos foram identificados um grande número de moléculas do sistema imune inato, como integrina beta, lectina tipo C, antígeno CD209 e outros. Também foram encontrados HSPs e genes relacionados com vias de doenças humanas, como "*Pathogenic Escherichia coli infection*", "*Vibrio cholerae infection*" e "*Legionellosis*". Os resultados levaram os autores a concluir que a exposição ao benzo(a)pireno inibe a performance do sistema imune inato do mexilhão, podendo induzir doenças infecciosas no animal.

Capítulo 2

Montagem e anotação funcional do transcritoma

2.1 Metodologia

2.1.1 Exposição dos mexilhões

Mexilhões adultos da espécie *Perna perna* (7-9cm) foram coletados na Fazenda de Aquicultura da Universidade Federal de Santa Catarina (UFSC) em Florianópolis (SC) e transferidos para o Laboratório de Bioensaios do Instituto do Mar da UNIFESP, em Santos (SP), onde foram acondicionados em 1 aquário aerado de 40L com água do mar filtrada e alimentados 1 vez ao dia com fitoplâncton comercial, e a água foi trocada 1 vez por dia. Os mexilhões foram aclimatados nessas condições por 6 dias antes da exposição, onde depois foram transferidos para 8 aquários aerados de 4L com água do mar, sendo 3 organismos por aquário.

A solução de ANT foi preparada dissolvendo o hidrocarboneto em DMSO até atingir uma concentração máxima de 1 mg/L, sem que a concentração de DMSO ultrapassasse 0,01% (v/v). A concentração foi escolhida com base em estudos prévios [1, 35]. A solução de ANT foi adicionada em 4 dos 8 aquários, caracterizando esse grupo como exposto. Nos demais foi adicionado apenas o DMSO até atingir a concentração máxima de 0,01% (v/v), caracterizando esse grupo como controle. O período de exposição foi de 24h, onde durante esse período e 48h horas antes da exposição, os mexilhões não foram alimentados para evitar a bioacumulação do ANT através da ração.

2.1.2 Extração do RNA e sequenciamento

O RNA total foi extraído das brânquias do animal, porque esse tecido tem um papel importante no processo de biotransformação [1], pois é um dos primeiros tecidos a ter contato com o ambiente contaminado, responsável por filtrar a água, e possivelmente o primeiro a sofrer alterações na expressão de genes. Leung et al. [15] observaram que genes relacionados com biodegradação e metabolismo de xenobióticos foram mais expressos nas brânquias e hepatopâncreas do mexilhão *Perna viridis*.

Após a exposição, os mexilhões foram mortos pelo rompimento do músculo adutor, e as brânquias foram dissecadas e imediatamente levadas para solução de extração do RNA, totalizando 24 amostras. A extração do RNA total das brânquias foi realizada com protocolo RNeasy mini kit (QIAGEN) seguindo as instruções do fabricante. Foi feito um

pool de amostras de RNA total para cada aquário ($n=3$) com o objetivo de diminuir a variabilidade da expressão dos genes entre os indivíduos, sendo, portanto, 8 amostras de RNA total.

Todo o procedimento de quantificação e análise de integridade das amostras, preparo das bibliotecas de cDNA e o sequenciamento foi realizado pelo Centro de Genômica Funcional localizado no Laboratório Multiusuário Centralizado da ESALQ. A quantificação e controle de qualidade da extração foi feito com Nanodrop 2000 (Thermo Fisher Scientific) e 2100 Bioanalyzer (Agilent). As bibliotecas de cDNA *stranded* foram preparadas com o protocolo TruSeq Stranded mRNA Sample Prep LS, seguindo instruções do fabricante, e sequenciadas em um único *lane* da plataforma Illumina HiSeq 2500 Flow Cell v4 utilizando o kit HiSeq SBS 4, gerando *reads* pareadas de 100 pb.

2.1.3 Análise química do tecido

A análise química foi realizada pela Dra. Márcia Caruso Bicego, Dra. Satie Taniguchi e Dr. Silvio Tarou Sasaki do Instituto Oceanográfico da USP (IO-USP). As amostras de tecido foram congeladas a seco e extraídas com Soxhlet com uma mistura de 50% de resíduo de n-hexano e diclorometano por 8 horas, baseado no procedimento descrito em [36]. Antes da extração, uma mistura de HPAs deuterados (d8-naftaleno, d10-acenafteno, d10-fenantreno, d12-criseno, and d12-perileno) foi adicionada a todas as amostras, os brancos e o material de referência do *National Institute of Standards and Technology - NIST (SRM 2974a - Organics in Freeze-Dried Mussel Tissue - Mytilus edulis)* como compostos substitutos. A coluna cromatográfica de gel-alumina de sílica (5% desativado) foi utilizada para limpar os extratos. A fração do tecido foi purificada por cromatografia líquida de alta performance (HPLC) para remover lipídeos. ANT foi analisado quantitativamente por cromatografia gasosa acoplada a espectrômetro de massas (GC/MS) da Agilent 6890 (GC) e 5973N (MS), no modo de íons selecionados (SIM).

2.1.4 Reconstrução da filogenia do mexilhão

Com intuito de conhecer o quão distante a espécie *P. perna* se encontra com relação a outros bivalves, foi feita a reconstrução da filogenia do mexilhão. As sequências do gene 18S rRNA foram colhidas do GenBank (Tabela 2.1) alinhadas com a ferramenta web MAFFT (26 de Jun, 2017) [37] e então o alinhamento foi verificado com GUIDANCE (26 de Jun, 2017) [38], ambos os *softwares* foram executados com os parâmetros padrão. A árvore foi gerada com o RAxML (v. 8.2.8) [39] utilizando os parâmetros "-m GTR-GAMMA -p 12345 -f a -N 1000 -x 12345". A visualização da árvore foi feita com o FigTree (v. 1.4.3) [40]. O grupo externo é composto por gastrópodes, com exceção da espécie *Katharina tunicata*, que é um poliplacóforo

2.1.5 Montagem *de novo* do transcrito

Foi verificada a qualidade do sequenciamento com FastQC (v. 0.10.1) [41]. As sequências com adaptadores do Illumina, bases de baixa qualidade ($<Q20$) e reads com alto teor de bases desconhecidas (N) foram "aparadas" com o Trimmomatic (v. 0.36) [42]. Se após a remoção de regiões da sequência, a mesma ficou com tamanho inferior a 30 pb, toda a sequência foi removida. Os parâmetros utilizados foram "*ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:7:1:true*", "*HEADCROP:13*", "*LEADING:30*", "*TRAILING:30*", "*MAXINFO:40:0.05*" and "*MINLEN:30*"

Tabela 2.1: Números de acesso das sequências do gene 18S rRNA depositadas no GenBank utilizadas para reconstrução da filogenia

Espécie	Número de acesso
Grupo externo	
<i>Aplysia californica</i>	AY039804.1
<i>Lottia gigantea</i>	FJ977632.1
<i>Haliotis tuberculata</i>	AF120511.1
<i>Katharina tunicata</i>	AY377650.1
Grupo interno	
<i>Crassostrea gigas</i>	AB064942.1
<i>Crassostrea virginica</i>	EU660792.1
<i>Ostrea edulis</i>	L49052.1
<i>Mytilus galloprovincialis</i>	L33452.1
<i>Mytilus edulis</i>	AY527062.1
<i>Perna perna</i>	DQ640517.1
<i>Perna viridis</i>	EF613234.1
<i>Pinctada margaritifera</i>	AJ389638.1
<i>Pinctada maxima</i>	AB214450.1

A montagem do transcrito então foi feita com as sequências limpas, utilizando Trinity (v. 2.0.2) [24] com o método *de novo*. Os parâmetros utilizados foram os padrões, com exceção do parâmetro `--SS_lib_type` que foi mudado para "RF" para considerar a orientação das *reads*. A redundância da montagem foi reduzida com CD-HIT-EST (v. 4.6.1) [43] estabelecendo um limite de 95% de similaridade.

O Trinity é dividido em 3 módulos: Inchworm, responsável por unir os reads em contigs lineares, fazendo alinhamentos a partir dos k-meros ($k = 25$ pb); Chrysalis, que faz a clusterização de *contigs* relacionados e então constrói uma estrutura de grafos de Bruijn onde os nós são os k-meros e as arestas são as sobreposições de tamanho $k-1$ entre os nós; e Butterfly, que analisa os caminhos formados pelo grafo de Bruijn com base nos reads e seus pares, identificando isoformas geradas a partir de eventos de splicing alternativo e transcritos derivados de genes parálogos, devolvendo ao final do processo todos os transcritos mais plausíveis [24].

A qualidade da montagem foi analisada com *scripts* do próprio Trinity e com o *software* BUSCO (v. 2.0.1) [44], que analisa a integridade dos transcritos montados, utilizando a linhagem "metazoa" e o modo "tran" como parâmetros. As sequências de aminoácidos foram preditas com o TransDecoder (v 2.0.1) [45] com ORFs de tamanho mínimo de 60 pb. Também foi feita uma curva de rarefação para saber se o sequenciamento foi suficiente para esgotar o universo de mRNAs do mexilhão. Para isso, foi feito o alinhamento de todas as *reads* com os transcritos montados utilizando o Bowtie2 (v. 2.2.6) [46] no modo de alinhamento local com parâmetros padrão. Apenas as *reads* que se alinharam uma única vez foram consideradas. Toda a etapa de verificação da qualidade da montagem que acabou de ser descrita foi feita após a etapa de descontaminação, que será descrita no item a seguir.

2.1.6 Descontaminação dos dados

O procedimento de descontaminação foi feito tanto a nível das *reads* quanto a nível dos *contigs*, levando em conta 3 possíveis fontes de contaminação: humano, devido a manipulação das amostras; algas, devido a ração; e bactérias.

Para humanos e algas, foi feito o alinhamento de todas as *reads* com o genoma humano (GRCh38.p10) e os genomas de espécies de algas presentes na ração (*Arthrospira platensis*, *Chlorella vulgaris*, *Nannochloropsis gaditana*, *Phaeodactylum tricornutum*) disponíveis no *Genome database* do NCBI (11 de Fev, 2017), utilizando Bowtie2 com parâmetros padrão. Todas as *reads* que alinharam uma única vez com algum dos genomas foram removidas. Para os *contigs* foi feita uma busca com BLASTn (v. 2.2.28+) [47] usando como referência os genomas humano e de algas, utilizando os parâmetros "*e-value*: 10^{-5} ; *identity*: 90%, *cover*: 70%". Os *contigs* que obtiveram hits foram removidos.

Para bactérias, as *reads* restantes da etapa anterior foram analisadas com um classificador metagenômico, CLARK (v. 1.2.3) [48] usando parâmetros padrão. As *reads* que foram classificadas foram alinhadas com os *contigs* restantes da etapa anterior usando Bowtie2 com parâmetros padrão. Os *contigs* que foram recuperados com o alinhamento, assim como as *reads* classificadas pelo CLARK foram removidas.

2.1.7 Anotação funcional

A anotação funcional do transcrito da *P. perna* foi feita com USEARCH [49] contra o banco NR do NCBI, utilizando o modo UBLAST e *e-value thresholds* de 10^{-5} e *bit score* de 50. Também foi feita a anotação utilizando BLASTn e tBLASTx contra os transcritos, e tBLASTn contra as proteínas da ostra *Crassostrea gigas* depositadas no *Genome database* do NCBI (24 de Fev, 2017), o bivalve com genoma completo mais próximo do *P. perna*.

Foi realizada uma busca por domínios conservados de famílias de proteínas utilizando Interproscan 5 [50] para anotar os transcritos montados, e o HMMER (v. 3.1.b1) [51] para anotar as sequências de aminoácido preditas. Ambos os *softwares* utilizaram a base de dados do Pfam (23 de Fev, 2017) usando como parâmetro *e-value* menor que 10^{-5} . Para cada domínio encontrado foi verificado quais termos do GO estavam associados.

Com o KAAS (*KEGG Automatic Annotation Server*) [52] foi verificado as enzimas, e as vias que elas participam, presentes entre os polipeptídeos preditos. Como parâmetros foram utilizados o conjunto de dados de eucariotos, adicionado os organismos *Crassostrea gigas*, *Lottia gigantea* e *Octopus bimaculoides*, e o método *single-directional best-hit* (27 de Set, 2017). Para recuperar a descrição das vias foi utilizado o pacote KEGGREST (v. 1.14) [53] do R.

2.2 Resultados e Discussões

2.2.1 Filogenia do *P. perna*

A partir da filogenia reconstruída (Figura 2.1) é possível afirmar que o gene 18S rRNA foi um bom marcador molecular, pois ele separou as espécies seguindo suas respectivas famílias, e em geral os clados estão bem suportados, considerando que quase todos os nós tiveram valores de suporte maiores que 70. Murgarella et al. [54] realizou o primeiro estudo com sequenciamento do genoma do *Mytilus galloprovincialis*. O gênero *Mytilus* é o mais próximo do *P. perna*, e, portanto, seria uma melhor referência, porém o genoma do

M. galloprovincialis ainda não está completo e está pouco anotado, o que faz a *C. gigas* a melhor referência para *P. perna*.

Figura 2.1: Reconstrução da filogenia do *P. perna* utilizando o gene 18S rRNA. Os números de suporte dos nós representam os valores de *bootstrap*, e a escala representa o número de substituições de nucleotídeo por sítio. As cores dos clados representam as famílias: vermelho = *Mytilidae*, azul = *Pteriidae*, Cinza = *Ostreidae*. O clado em preto representa o grupo externo

2.2.2 Concentração de ANT no tecido mole

A concentração de ANT foi medida no tecido mole do mexilhão para verificar se o animal foi capaz de absorver ANT da água. Todas as amostras do grupo exposto apresentaram concentrações mais elevadas que as do grupo controle (Tabela 2.2), o que demonstra que a exposição foi realizada com sucesso.

Tabela 2.2: Concentração de ANT (ng/g) no tecido mole do *P. perna* após 24h de exposição.

Grupo	Amostra 1	Amostra 2	Amostra 3	Amostra 4
Controle	3,18	8,9	18,9	8,9
Exposto	2.716	5.009	5.062	3.087

2.2.3 Sequenciamento e montagem do transcrito

O sequenciamento das 8 bibliotecas de cDNA do mexilhão gerou, aproximadamente, entre 33 - 37 milhões de pares de *reads*. Após as etapas de filtração e descontaminação restaram 32,5 - 36 milhões de pares de *reads* com *phred score* maior que Q20. Todas as *reads* foram utilizadas na etapa de montagem dos transcritos com o Trinity mais CD-HIT-EST, gerando 253.267 *contigs*, em que após a etapa de descontaminação restaram 231.728 (91,5%), com tamanho médio de 720 pb, variando entre 224 - 31327 pb (Figura ?), e N50 de 1083 pb (Tabela 2.3). Os relatórios gerados pelo FastQC se encontram no Anexo 1.

Figura 2.2: Distribuição dos tamanhos dos *contigs* montados. O gráfico a) apresenta todos os tamanhos, enquanto o gráfico b) é um recorte do gráfico a) limitando o tamanho máximo em 6.000 pb.

Tabela 2.3: Estatísticas gerais sobre a montagem *de novo* do transcrito do *P. perna*. As amostras C1, C2, C3 e C4 representam as réplicas do grupo controle, enquanto as amostras A1, A2, A3 e A4 representam as réplicas do grupo exposto.

Estatísticas gerais	C1	C2	C3	C4	A1	A2	A3	A4
Total bruto de <i>reads</i> pareadas	35.297.747	32.754.613	32.911.285	33.699.225	33.783.842	35.117.884	36.891.020	35.592.867
Total de <i>reads</i> pareadas após filtro	35.059.465	32.564.598	32.710.634	33.468.007	33.539.601	34.753.917	35.748.586	35.307.582
Total de <i>unigenes</i>				187.804				
Total de transcritos				231.728				
Tamanho médio (pb)				720				
Tamanho mínimo (pb)				224				
Tamanho máximo (pb)				31.327				
Tamanho < 500 pb				141.442				
Tamanho $\geq$ 500 pb e < 5.000 pb				88.534				
Tamanho $\geq$ 5.000 pb				1.753				
N50 (pb)				1.083				

A montagem dos transcritomas dos mexilhões *Perna viridis* [15, 34] e *Mytilus galloprovincialis* [32] também foram realizadas com o Trinity, e o resultado obtido foram transcritos com tamanho médio de 1.264, 981 e 570 pb, respectivamente, e N50 de 2.868, 2.383 e 774 pb, respectivamente. Comparando os estudos citados anteriormente com os resultados obtidos no presente trabalho é possível dizer que a qualidade da montagem se encontra dentro de padrões aceitáveis. A análise de integridade da montagem do transcritoma revelou 803 grupos ortólogos completos e cópia-única em um total de 978, o que indica que dos grupos de genes ortólogos que se espera encontrar em uma espécie da linhagem metazoa, 82,1% foi encontrado na montagem realizada. Dos grupos ortólogos restantes, 167 (17%) foram encontrados duplicados e 8 (0,8%) estavam fragmentados (Tabela 2.4). Não foi reportada a ausência de nenhum grupo ortólogo. Os resultados obtidos pela análise indicam que grande parte da montagem do transcritoma está íntegra, podendo afirmar que ela foi bem sucedida.

Tabela 2.4: Relatório da análise de integridade da montagem do transcritoma do *P. perna* realizada com o BUSCO.

Grupos ortólogos	#
Cópia-única	803(82,1%)
Duplicado	167 (17%)
Fragmentado	8 (0,8%)
Ausente	0 (0,0%)

Observando a curva de rarefação a seguir (Figura 2.3) é possível notar que com aproximadamente metade das *reads* sequenciadas (145 milhões, 53%) foi possível recuperar 209.832 (90,5%) dos transcritos montados, e a curva está quase atingindo o platô. Ambas constatações indicam que a profundidade do sequenciamento foi satisfatória, considerando que o presente trabalho é a primeira tentativa de sequenciamento e análise do transcritoma da espécie *P. perna*.

Figura 2.3: Curva de rarefação feita a partir do alinhamento das *reads* com o transcritoma montado

2.2.4 Anotação funcional do transcrito

Inicialmente a anotação funcional do transcrito foi realizada utilizando o *nr database* do NCBI, resultando em 53.753 (23,2%) *contigs* que tiveram pelo menos um *hit* (Tabela 2.5). O número de *contigs* anotados foi maior que o apresentado por outros estudos com mexilhões (41.686 em *P. viridis* [15], 26.039 em embriões de *P. viridis* [34] e 45.182 em *M. galloprovincialis* [32]). A maior parte das sequências anotadas foram similares a sequências da ostra *C. gigas*, assim como observado para *P. viridis* [15].

Tabela 2.5: Anotação funcional dos *contigs*. Pfam (nucleotídeo) são resultados obtidos a partir do Interproscan onde o arquivo de entrada continha sequências de nucleotídeos. Pfam (aminoácido) são resultados obtidos a partir do HMMER onde o arquivo de entrada continha sequências de aminoácidos

Base de dados	# de <i>contigs</i>
NR NCBI	53.753
<i>C. gigas</i> genome	54.884
Pfam (nucleotídeo)	25.893
GO (nucleotídeo)	14.786
Pfam (aminoácido)	40.982
GO (aminoácido)	26.028
KEGG	26.298
Total de <i>contigs</i> anotados (não-redundante)	66.563

Também foi feita a anotação usando os transcritos e proteínas do genoma da *C. gigas* depositadas no *Genome database* do NCBI como referência. Apenas 283 (0,12%) dos *contigs* foram anotados a nível de nucleotídeo, o que era esperado, porque o mesmo foi observado para ostra *Saccostrea glomerata* [55]. No entanto, a nível de aminoácido foram anotados 54.884 (23,7%) *contigs* (Tabela 2.5), recuperando 18.486 (36,3%) transcritos e 18.612 (39,8%) proteínas de *C. gigas*.

A busca por domínios conservados de proteína a nível de nucleotídeo resultou em 25.893 (11,2%) *contigs* com pelo menos um domínio encontrado, enquanto a nível de aminoácido foram encontrados 40.982 (17,6%) *contigs* com pelo menos um domínio. Dos *contigs* encontrados com domínios conservados de família de proteínas, 14.786 (57,1%) e 26.028 (63,5%) a nível de nucleotídeo e aminoácido, respectivamente, tinha pelo menos um termo do GO atribuído (Tabela 2.5). O banco de dados do Gene Ontology é dividido em três classes de ontologia: *biological process* (BP), que se refere a um objetivo biológico na qual um gene ou um produto gênico contribui para alcançá-lo; *molecular function* (MF), que define a atividade bioquímica de um produto gênico; e *cellular component* (CC), que se refere ao local na célula onde o produto gênico é ativo [56].

Processo de oxidação e redução é o processo biológico de maior abundância de trans-

critos associados com essa função na brânquia (Figura 2.4), e ele está relacionado com várias funções comuns na célula (respiração celular, glicólise...), mas também está relacionado com atividades de biotransformação de xenobióticos e antioxidantes, uma vez que genes como CYP, SOD, GPx e CAT são atribuídos à essa categoria do GO. Trabalhos já mostraram que genes de CYP são expressos em brânquias de outros bivalves na presença de xenobióticos [1, 15], e também existem trabalhos que mediram a atividade de enzimas antioxidantes em brânquias de *P. perna* e como elas respondem na presença de contaminantes [10, 13, 14] ou quando expostos ao ar [57].

Figura 2.4: Os 20 termos GO mais abundantes na categoria BP a nível de aminoácido. O eixo X representa o número de transcritos associados

A abundância de transcritos associados a função de transporte de membrana pode ser uma evidência que as brânquias podem exercer um papel na osmorregulação. Sabe-se que as brânquias de bivalves marinhos podem absorver aminoácidos livres (FAA), possivelmente para manutenção da concentração de FAA na célula para compensar variações na concentração iônica do meio externo, mantendo o volume celular [58], lembrando que bivalves marinhos, na sua grande maioria, são osmoconformadores [59], ou seja, a osmolaridade interna do organismo tende a acompanhar a osmolaridade do meio em que vive. No entanto, Rola et al [59] acreditam que mexilhões *P. perna* podem adotar mecanismos de osmorregulação dependendo de sua origem e tempo de exposição ao estresse osmótico, isso porque eles observaram que mexilhões estuarinos (*P. perna*) retornaram aos valores iniciais de osmolaridade e concentração iônica depois de 4 dias em meio hipoosmótico, e paralelamente a atividade da enzima Na^+/K^+ ATPase aumentou. O termo "ion transport" do GO está entre os 20 termos GO mais abundantes em BP (Figura 2.4), o que é coerente com a hipótese acima. Apenas em brânquias de crustáceos e peixes se conhece sua participação na osmorregulação, enquanto em bivalves essa função tem sido pouco estudada [32].

Existem também transcritos associados com a função de sinalização que estão abundantes nas brânquias, o que é coerente com o contexto do tecido, que se encontra em contato direto com o meio, cercado de estressores bióticos e abióticos, e por isso todas as mudanças captadas devem ser sinalizadas para as células para que haja a adaptação apropriada do organismo. Moreira et al [32] também observou que transcritos associados com a função de sinalização estavam abundantes em brânquias do mexilhão *Mytilus galloprovincialis*. Na categoria MF, o termo "protein binding" foi o mais abundante (Figura 2.5), e esse termo está relacionado a diversos processos biológicos, incluindo genes de defesa [60]. As demais figuras com resultados da anotação com GO estão na seção Anexo 2.

Figura 2.5: Os 20 termos GO mais abundantes na categoria MF a nível de aminoácido. O eixo X representa o número de transcritos associados

A anotação com o banco de vias do KEGG resultou em 26.298 (11,3%) de transcritos anotados, resultado superior ao obtido para o estudo com o *P. viridis* [15]. As vias com maior número de transcritos associados estão relacionadas com transdução de sinal, doenças infecciosas, câncer, sistema imune e endócrino (Figura 2.6). Moreira et al [32] observaram que vias relacionadas com transdução de sinal, doenças infecciosas e câncer também estavam abundantes em transcritomas de *Mytilus galloprovincialis* e *Crassostrea gigas*.

Os resultados obtidos com anotação com GO e KEGG sugerem que brânquias podem ter um papel em sinalização e resposta do sistema imune. A provável explicação foi dada por Zhang et al. [60] que diz que ostras, e provavelmente outros moluscos, adaptaram-se para lidar com estressores ambientais durante a evolução, porque no estudo foi feita a comparação do genoma da ostra *C. gigas* com outros 7 genomas (*Nematostella vectensis*, *Strongylocentrotus purpuratus*, *Ciona intestinalis*, *C. elegans*, *D. melanogaster*, *Danio rerio* and *Homo sapiens*), revelando 8.654 genes que são exclusivos da ostra, sendo na sua

grande maioria associados com termos GO relacionados com defesa do hospedeiro, como "*protein binding*", "*apoptosis*", "*cytokine activity*" e "*inflammatory response*".

Figura 2.6: Anotação das proteínas, preditas a partir das sequências de transcritos, com o banco de vias do KEGG.

Sistema imune inato

Moluscos não possuem sistema imune adaptativo como os vertebrados, e por isso dependem do sistema imune inato, também conhecido como sistema imune não-específico, para se proteger de patógenos [61]. Recentes estudos de genômica em ostras mostram uma diversificação em receptores e efetores do sistema imune inato, provavelmente para

compensar a falta de um sistema adaptativo [61].

No transcritoma do *P. perna*, 3476 transcritos foram atribuídos a vias do KEGG de sistema imune (Figura 2.6). Entre eles a grande maioria são vias de sistema imune inato, como as vias de sinalização de Toll e Imd, a com maior número de transcritos associados, que ativam a produção de peptídeos antimicrobianos que detectam fungos, bactérias Gram-positiva e Gram-negativas e fatores de virulência; receptores do tipo NOD, que são receptores intracelulares que detectam padrões moleculares associados a patógenos (MAMP) de longo espectro [62], sendo essa via a mais representativa (Figura 2.7); receptores do tipo Toll (TLR), que são receptores ligados a membrana que fazem reconhecimento de MAMPs; receptores do tipo RIG I, que são helicases citosólicas de RNA que reconhecem produtos de replicação viral intracelular; e interleucinas-17, que são uma família de citocinas que atuam tanto na resposta inflamatória aguda quanto crônica [63] (Figura 2.7).

Figura 2.7: As vias de sistema imune anotadas com o banco de dados KEGG.

Existem também transcritos com domínios de proteínas do sistema imune inato que não foram identificados nas bases do KEGG, o que aumenta ainda mais o número de transcritos que provavelmente exercem essa função em *P. perna*. Entre eles estão mais TLRs, lectinas do tipo C, que são proteínas de reconhecimento de carboidrato que atuam no sistema imune inato em invertebrados, fazendo reconhecimento de patógenos [34], fibrinogênios, que são lectinas com pelo menos um domínio fibrinogênio, onde sua função dentro do sistema imune inato ainda não é bem conhecida [61] e proteínas contendo domínios de C1q (C1qDC), que detectam MAMPs e ativam fagocitose [64]. A figura 2.8 apresenta uma comparação entre o número de genes de sistema imune inato entre espécies modelo [61] com alguns dos genes encontrados no *P. perna*.

Os resultados obtidos no presente estudo vão ao encontro com o raciocínio seguido por outros estudos [60, 61], em que moluscos expandiram o número de genes do sistema imune

Figura 2.8: Genes do sistema imune inato encontrados em *P. perna* comparando com alguns organismos modelos. O número de genes das espécies modelo foram retirados do artigo do Guo et al [59]. Os números para *P. perna* representam os *unigenes* montados pelo Trinity, e portanto, esses valores são aproximações do número real de genes que se espera encontrar.

inato como uma forma de adaptação aos estressores ambientais, já que invertebrados não possuem um sistema imune adaptativo [60, 61].

Capítulo 3

Genes de biotransformação de HPAs e análise de expressão diferencial

3.1 Metodologia

3.1.1 Estimação dos níveis de expressão gênica

O nível de expressão foi estimado a nível de isoforma, utilizando os softwares Kallisto (v. 0.42.5) [27] e Salmon (v. 0.6.1) [28] com parâmetros padrão, exceto pela orientação das *reads* que foi modificado para opção correspondente à orientação "RF" do Trinity. Transcritos que tiveram a contagem estimada menor que 5 na somatória de todas as bibliotecas foram descartados.

Ambos os softwares implementam métodos que não necessitam da etapa de alinhamento das *reads* com os transcritos, que podem introduzir alguns vieses nas estimativas. Comparando os resultados de ambos os *softwares* e verificando quais dados são concordantes, é possível diminuir o viés da estimativa.

3.1.2 Cálculo da expressão diferencial

A análise de expressão diferencial foi realizada com o *software* EdgeR (v. 3.14.0) [65], sendo o *threshold* do FDR (*false discovery rate*) menor ou igual a 0,05 e LFC (*Log₂ fold change*) maior ou igual ao valor absoluto de 1. A comparação dos resultados entre os *softwares* Kallisto e Salmon foi feita com diagramas de Venn, que foram construídos com o pacote do R, VennDiagram (v. 1.6.17).

3.2 Resultados e Discussões

3.2.1 Transcritos relacionados com biotransformação de HPA

Ecossistemas marinhos e costeiros são suscetíveis a contaminação por uma diversidade de xenobióticos, incluindo HPA's, devido a intensa atividade humana nessas regiões. Portanto, seres vivos que habitam esses ecossistemas precisam de mecanismos de detoxificação contra esses contaminantes para sobreviver. No transcrito do *P. perna* foram identificados 209 transcritos envolvidos com a via de biodegradação de xenobióticos e metabolismo (Figura 2.6), existindo também aqueles que foram anotados por outros bancos de dados (NR NCBI, Pfam e GO). Entre esses foram identificados genes de biotransformação de fase I, II e III. Esses resultados já representam um avanço para a ecotoxicologia, visto que, para conhecimento do presente autor, este é o primeiro estudo que identificou transcritos possivelmente envolvidos com metabolismo de xenobióticos para essa espécie.

Para HPAs, a via de sinalização para ativar esses genes é mediada pelo receptor de aril hidrocarboneto (AhR) [66, 67, 68]. Sabe-se que AhR regula a ativação transcricional de algumas isoformas de citocromo P450 (CYP) [66, 67] e glutathione S-transferase (GST) [69, 70], tendo ambas um importante papel nas fases I e II da biotransformação de HPAs em bivalves [1].

Citocromo P450

Enzimas CYP mediam reações da fase I de biotransformação de substâncias exógenas, como drogas, HPAs, pesticidas e outros xenobióticos lipofílicos, mas também são importantes para metabolismo de substâncias endógenas [71, 72]. CYPs são monooxigenases que catalisam reações de oxidação de substâncias lipofílicas, transferindo um átomo de oxigênio para a substância, aumentando sua polaridade.

O transcrito do *P. perna* revelou 124 transcritos similares a 31 tipos de CYP de 10 famílias e 18 subfamílias pertencentes aos clãs 2, 3, 4 e mitocondrial. Esses resultados foram comparados com estudos de outros bivalves. Esses resultados foram comparados com outros estudos na literatura [32, 15] e com dados de transcritos de *C. gigas* depositados no Genome database do NCBI (Figura 3.1, 3.2).

Figura 3.1: Comparações dos números absolutos de transcritos similares a membros da família CYP identificados em *P. perna* com outros bivalves.

Figura 3.2: Comparações dos números de transcritos similares a membros da família CYP identificados em *P. perna* com outros bivalves, normalizados pelo total de transcritos identificados na espécie.

Observando a comparação dos números absolutos de transcritos similares a CYPs vemos que o *P. perna* tem uma quantidade similar a outros bivalves, levando em conta que para o *P. perna* são dados gerados apenas das brânquias, enquanto os outros bivalves envolvem mais de um tecido. Porém, ao se deparar com os números normalizados, percebe-se que a *C. gigas* tem uma quantidade muito maior que os demais bivalves, isso porque o genoma da ostra já está completo e o número total de transcritos já é conhecido (53.160 transcritos, 03 de agosto de 2017), enquanto os demais bivalves representam os primeiros estudos de transcriptoma, e o número de transcritos totais ultrapassam os valores de 150 mil transcritos, sendo muitos deles artefatos dos *softwares* de montagem.

Alguns trabalhos mostram que a família CYP 1 (CYP 1A1, 1A2 e 1B1) está relacionada com o processo de biotransformação de HPAs em humanos [67] e peixes [66], enquanto genes da família CYP 2 foram regulados positivamente quando a ostra *Crassostrea brasiliiana* foi exposta a fenantreno (PHE) [1]. A isoforma CYP2-like 2 da *C. brasiliiana* é muito próxima da CYP 2C8 da *C. gigas* [1]. No presente estudo foram encontrados transcritos similares a isoformas CYP 1A2, 1B1 e 2C8 de *C. gigas*.

Visualmente não é possível identificar uma diferença significativa entre as isoformas de CYP, com exceção da isoforma 27C1, e em alguns casos a média do grupo controle é superior ao grupo exposto (Figura 3.3), o que nos faz pensar que esses são os níveis basais de expressão desses genes, ou algum fator, por exemplo o DMSO, presente em ambos os grupos ativou a expressão desses genes. Mas aqui ainda não está sendo feita uma análise de expressão diferencial e não se está considerando os dados gerados pelo outro software utilizado (Salmon), e por isso não se pode afirmar se há ou não diferenças significativas de expressão gênica entre os grupos.

Figura 3.3: Níveis de expressão, quantificados em TPM pelo Kallisto, das 10 isoformas de CYPs mais expressas identificadas em *P. perna* em ambos os grupos.

Glutathione S-transferase

Enzimas Glutathione S-transferase tem um importante papel nas reações de fase II de biotransformação, catalizando a conjugação da glutathione reduzida (GSH) a um composto hidrofóbico [15, 68, 73, 74]. No presente estudo, 63 sequências foram similares a 8 famílias de GSTs citosólicas (*alpha*, *kappa*, *omega*, *mu*, *pi*, *sigma*, *tau* e *theta*), além de famílias microsossomais e sequências com domínios de GST. Esses resultados foram comparados com outros estudos na literatura [32, 15] e com dados de transcritos de *C. gigas* depositados no Genome database do NCBI (Figura 3.4, 3.5). Pode-se dizer que o mesmo padrão observado para CYP é observado para GST.

Figura 3.4: Comparações dos números absolutos de GSTs identificadas em *P. perna* com outros bivalves.

Boutet et al [73] expôs *C. gigas* a uma mistura de hidrocarbonetos e notou que as isoformas *mu*, *omega* e *pi* de GST foram reguladas positivamente. Lünchmann et al. [1]

Figura 3.5: Comparações dos números absolutos de GSTs identificadas em *P. perna* com outros bivalves, normalizados pelo total de transcritos identificados na espécie.

observou que as isoformas GST *omega-like* e *microsomal-like* foram reguladas positivamente em *C. brasiliensis* quando exposta a PHE, e essas isoformas eram próximas das isoformas GST *omega-1* e *microsomal-3*, respectivamente, da *C. gigas*. As GSTs *omega-1* e *microsomal-3* de *C. gigas* também foram similares a algumas sequências identificadas no *P. perna*, e elas estavam entre as 10 isoformas mais expressas. Quanto ao nível de expressão, o mesmo observado com CYPs foi observado para GSTs (Figura 3.6).

Figura 3.6: Níveis de expressão, quantificadas em TPM pelo Kallisto, das 10 isoformas de GSTs mais expressas identificadas em *P. perna* em ambos os grupos.

3.2.2 Análise de expressão diferencial

Após estimar os níveis de expressão dos transcritos, foi verificado se os dados seguiam uma distribuição binomial negativa, visto que o EdgeR assume que os dados seguem esse modelo de distribuição. Para isso foi feito um gráfico que relaciona a média e a variância dos dados (Figura 3.7). Observa-se que os dados estão mais próximos da linha azul, que representa um modelo de distribuição binomial negativa, do que da linha preta, que representa um modelo de distribuição de Poisson, portanto, é possível o uso do pacote EdgeR para a análise de expressão diferencial dos transcritos nesse conjunto de dados.

Figura 3.7: Relação da média com a variância obtida da estimação dos níveis de expressão dos transcritos. A linha azul representa o modelo de distribuição binomial negativa, a linha preta o modelo de distribuição de Poisson, e os pontos cinzas representam os dados observados.

Foram identificados 353 e 301 TDEs a partir das estimativas do Salmon e Kallisto, respectivamente. A grande maioria dos TDEs (269 transcritos, 70%) foram identificados nos dois grupos de dados gerados (Figura 3.8), e entre eles, 77-78% não foram anotados, ou anotados como não-caracterizado ou hipotético (144 (77%) entre os regulados positivamente e 64 (78%) entre os regulados negativamente) (Tabela 3.1).

Um *volcano plot* foi feito para visualizar o efeito da exposição ao ANT na regulação da expressão gênica, representado pelo eixo X, e a significância estatística desses resultados, representado pelo eixo Y. Quanto mais próximos dos extremos do eixo X, maior a influência do contaminante na regulação da expressão daquele gene, e quanto mais próximo ao topo do gráfico, maior a sua significância (Figura 3.9).

Figura 3.8: Diagrama de Venn dos TDEs identificados a partir dos dados de estimativas do Salmon e Kallisto.

Figura 3.9: *Volcano plot* dos TDEs identificados a partir dos dados de estimativas do Salmon e Kallisto.

Tabela 3.1: Os 20 TDEs mais expressos/inibidos em *P. perna* após 24h de exposição ao ANT.

	Regulação positiva	Regulação negativa
1	Hypothetical protein	Interferon-inducible GTPase (IIGP)
2	Hypothetical protein	interferon-induced GTPase 1
3	titin	Hypothetical protein
4	Fibrinogen C domain	Immunoglobulin I-set domain
5	Lectin C domain	hemicentin-1
6	Lecithin retinol acyltransferase	Hypothetical protein
7	Immunoglobulin domain	serine/threonine protein kinase
8	Hypothetical protein	Hypothetical protein
9	Immunoglobulin domain	hemicentin-1
10	uncharacterized protein [<i>Crassostrea gigas</i>]	Hypothetical protein
11	E3 ubiquitin-protein ligase	Hypothetical protein
12	Aga1p [<i>S. cerevisiae</i>]	von Willebrand factor D and EGF domain
13	Hypothetical protein	Hypothetical protein
14	Hypothetical protein	Hypothetical protein
15	uncharacterized threonine-rich GPI-anchored	Hypothetical protein
16	mucin-2-like	Hsp70-Hsp90 organizing protein 1
17	Hypothetical protein	Hypothetical protein
18	Hypothetical protein	Hypothetical protein
19	Hypothetical protein	titin isoform X8
20	Inhibitor of Apoptosis domain	Hypothetical protein

Entre os TDEs anotados regulados positivamente foi identificado uma peroxidase (peroxiredoxin 1 [EC:1.11.1.15]). Peroxidases são enzimas com atividade antioxidante, e por isso, a ativação desses genes podem indicar que os mexilhões estavam sofrendo estresse

oxidativo, muito provavelmente causado pela exposição ao ANT. Não foi identificado nenhum CYP, GST ou outros genes relacionados com biotransformação de HPA regulados positivamente ou negativamente, o que não era esperado, já que a concentração de ANT estava 268 a 715 vezes maior no grupo exposto comparado ao controle (Tabela 2.2). Uma possível explicação seria a toxicidade do próprio solvente (DMSO), que foi suficiente para ativar os mecanismos de biotransformação em ambos os grupos. Porém outros trabalhos, como o de Lüchmann et al [1], em que o DMSO foi utilizado como solvente sob mesma concentração e foi observada a regulação positiva dos genes de biotransformação em ostras expostas a fenantreno. O tempo de exposição também pode ser a causa, sendo 24h insuficiente ou longo demais para identificar os genes de biotransformação, apesar de outros estudos utilizarem a mesma duração de exposição a contaminantes para outros bivalves [1, 15, 34], ou ainda, outro tecido e responsável por metabolizar o ANT, apenas de outros trabalhos mostrarem evidências do papel da brânquia na biotransformação de HPAs [1, 15].

Entre os TDEs regulados positivamente foram encontrados transcritos com domínios de: lectina tipo C; fibrinogênio; imunoglobulinas; e inibidor de apoptose, domínios esses que podem estar associados a proteínas que participam do sistema imune inato. Jiang et al [34] observaram alterações na resposta do sistema imune inato em embriões de *P. viridis* devido a contaminação por benzo(a)pireno, onde proteínas com domínio lectina tipo C também foram reguladas positivamente. Entre os regulados negativamente, os mais inibidos foram genes de GTPases induzidas por interferons (IIGP), que são citocinas envolvidas no combate a infecção microbiana intracelular, sendo essas responsáveis por sinalizar a ativação de genes que codificam efetores e reguladores de proteínas microbicidas [75].

Esses resultados sugerem que a exposição ao ANT poderia estar alterando a capacidade de resposta do sistema imune inato do *P. perna*, deixando-o mais vulnerável a infecções devido a inibição das IIGPs. Por outro lado, outros genes de resposta do sistema imune estariam sendo ativados, podendo atuar como "substitutos" para se defender de patógenos. O que não é possível saber é se o ANT provocou a ativação desses genes, ou se esses genes foram ativados devido a inibição dos IIGP provocada pelo ANT, como uma via alternativa.

Em humanos e camundongos, a ativação de genes do sistema imune após exposição a HPAs já é conhecida. Sabe-se que o receptor AhR modula a expressão de genes sistema imune inato e adaptativo [76, 77]. Também é observado a supressão da expressão de CYPs durante respostas inflamatórias, apesar desse mecanismo ainda não ser bem elucidado [76, 78]. Acredita-se que algumas citocinas sejam responsáveis pela inibição de CYPs, iniciada por um estímulo de estresse [78]. Uma hipótese seria que mecanismos similares aos observados para humanos e camundongos ocorra em *P. perna*, o que explicaria a ativação de genes do sistema imune inato e a inibição da expressão de CYPs. Uma segunda hipótese seria a questão levantada por Zhang et al [60] e Guo et al [61], que dizem que bivalves se adaptaram ao estresse ambiental ao longo da evolução, e por isso houve uma diversificação de genes relacionados com sistema imune inato e resposta ao estresse, sendo possível que alguns genes de sistema imune possam ter adquirido uma nova função, a qual seria de resposta a estressores abióticos [61]. Mas não se pode ignorar o fato de que quase 80% dos TDEs não tiveram função atribuída, sendo possível que esses genes possam ter algum papel na biotransformação de ANT, mas que seus mecanismos ainda são desconhecidos. Os demais TDEs que não estão relacionados com sistema imune ou biotransformação são necessários mais estudos para elucidar seus papéis na resposta a contaminação por ANT.

A não expressão diferencial de genes de biotransformação pode também ser devido a problemas com as réplicas biológicas utilizadas no estudo, que podem estar distantes uma das outras, superestimando a variabilidade biológica, dificultando na identificação dos TDEs, uma vez que um transcrito só é classificado como TDE se a diferença do nível de expressão entre os grupos for maior que a variância interna dos grupos. Um *MDS plot* (*Multi-dimensional scaling plot*) foi feito para visualizar a similaridade das réplicas entre si (Figura 3.10). Também foi feito uma análise de componentes principais (PCA) para cada grupo (Figura 3.11, 3.12).

Figura 3.10: MDS plot das réplicas utilizadas no estudo. A1, A2, A3 e A4 representam as réplicas do grupo exposto a ANT, enquanto C1, C2, C3 e C4 representam as réplicas do grupo controle.

Visualmente é possível observar que existe uma inconsistência das réplicas de cada grupo. As réplicas A1 e A3 são mais próximas, mas as demais (A2 e A4) são mais distantes. O mesmo ocorre para réplicas do grupo controle. As réplicas C1 e C4 são mais próximas entre si, mas C2 e C3 são distantes. Faz-se necessário uma análise de correlação para saber o quanto cada réplica se correlaciona entre as demais, e se essa correlação é significativa. Não pudemos realizar essa análise por falta de tempo hábil em relação à defesa desta dissertação. Entretanto pretendemos realizar esta análise, para que o artigo a ser submetido com os resultados deste trabalho leve em conta esses resultados. Se for constatado que as réplicas A2, A4, C2 e C3 não se correlacionam com as demais réplicas, faremos novas análises de expressão diferencial levando em conta apenas as réplicas mais correlacionadas, o que provavelmente terá mudanças na lista dos TDEs aqui apresentados.

Figura 3.11: PCA plot das réplicas do grupo exposto.

Figura 3.12: PCA plot das réplicas do grupo controle.

Conclusão

Para o presente trabalho foi proposto realizar a exposição de mexilhões marinhos (*P. perna*) ao ANT, e então analisar o transcrito desse animal na tentativa de identificar quais genes são ativados ou inibidos pela presença do contaminante.

Os resultados obtidos de métricas da montagem dos *contigs* (tamanho médio e N50) comparada a outros mexilhões, juntamente com análise de integridade dos transcritos e a curva de rarefação, mostram que a montagem *de novo* do transcrito foi realizada de maneira satisfatória, considerando ainda que se trata do primeiro estudo com RNA-Seq realizado para essa espécie.

A anotação funcional dos transcritos apontou 5 possíveis funções desempenhadas pelas brânquias desse mexilhão, além da troca gasosa e filtração, que são: biotransformação de xenobióticos, combate a estresse oxidativo, transdução de sinal do meio externo, resposta imunológica e osmorregulação. Todas essas funções já haviam sido observadas para outros bivalves, inclusive mexilhões de outras espécies.

Com relação a osmorregulação, os resultados aqui obtidos corroboram a hipótese formulada com base em dados experimentais em um estudo recente, que diz que possivelmente brânquias de mexilhões *P. perna* fazem osmorregulação sob estresse osmótico prolongado [59]. O papel da brânquia na regulação da resposta imunológica ficou ainda mais evidente com a exposição ao ANT. Genes relacionados com essa função foram tanto ativados como inibidos pela presença do HPA. Essas descobertas contribuem para o panorama geral da biologia dessa espécie de mexilhão, o que permite desenvolver novas aplicações em outras áreas da ciência, como a aquicultura e a ecotoxicologia.

Também foi possível identificar genes de biotransformação das 3 fases, e dentre eles estão CYPs e GSTs. Ambos os genes estavam sendo expressos no grupo exposto, mas por motivos não claros nenhum desses genes foi classificado como diferencialmente expresso. É possível que *P. perna* utilize mecanismos de biotransformação para sobreviver a contaminação por ANT que são desconhecidos, considerando que quase 80% dos TDEs não foram anotados pela metodologia implementada. Outra hipótese é a inibição da expressão desses genes por estímulos de estresse, como inflamações e espécies reativas de oxigênio. O tempo de exposição e o solvente utilizado também podem ser as causas. Outro fator é a qualidade das réplicas utilizadas, que, visualmente, não são consistentes, sendo necessário quantificar a correlação que existe entre as réplicas, e verificar se as correlações são significativas.

Assim como observado para a ostra *C. gigas*, *P. perna* também possui um diversificado número de genes de sistema imune inato, o que pode ser resultado de uma adaptação a estressores ambientais ao longo da evolução. Ao longo dessa diversificação é possível que alguns genes de sistema imune inato possam ter adquirido a função de combater estressores abióticos, o que pode ser um dos motivos da ativação de genes de sistema imune inato durante a exposição a ANT.

A ativação dos genes do sistema imune inato também pode ser devida à regulação do

receptor AhR que é ativado na presença de HPAs, o mesmo receptor que regula a expressão de CYPs e GSTs. Esse mecanismo já é conhecido para humanos e camundongos, e se sabe que o AhR regula a expressão de genes de sistema imune inato e adaptativo nessas duas espécies.

Como perspectivas futuras são propostas: realizar análises de correlação entre as réplicas e refazer análises de expressão diferencial apenas com as réplicas que se correlacionam com outras do mesmo grupo; fazer uma busca por RNAs não-codificadores longos entre os 165.165 transcritos que não foram anotados por nenhuma das ferramentas utilizadas e verificar se eles respondem a presença de HPAs; desenhar *primers* para alguns dos TDEs para validação por qPCR, e uma vez validados, fazer coletas de campo para testar a eficiência dos *primers* e propor novos biomarcadores de exposição para *P. perna* que respondam a presença de ANT; sequenciar o genoma da espécie e remontar os transcritos utilizando o genoma como referência; realizar novas exposições com diferentes concentrações de ANT, diferentes tempos de exposição e três grupos: exposto à ANT (ANT + DMSO + água do mar), exposto ao solvente (DMSO + água do mar) e controle (água do mar apenas)

Em resumo, as contribuições deste trabalho foram:

1. Elaboração de um pipeline computacional rigoroso para análise de dados de transcrito, com base em pipelines semelhantes da literatura;
2. Obtenção do primeiro transcrito do mexilhão *Perna perna*, o que gerou evidências das possíveis funções das brânquias;
3. Determinação dos genes diferencialmente expressos sob exposição ao antraceno, o que revelou um papel de genes do sistema imune

Este trabalho, no seu conjunto, é portanto uma contribuição para o melhor entendimento da biologia do mexilhão *Perna perna*.

Referências Bibliográficas

- [1] Karim H. Lüchmann, Alcir L. Dafre, Rafael Trevisan, John A. Craft, Xiang Meng, Jacó J. Mattos, Flávia L. Zacchi, Tarquin S. Dorrington, Declan C. Schroeder, and Afonso C.D. Bainy. A light in the darkness: New biotransformation genes, antioxidant parameters and tissue-specific responses in oysters exposed to phenanthrene. *Aquatic Toxicology*, 152:324–334, jul 2014.
- [2] Carlos Gravato, Joana R. Almeida, Carlos Silva, Cristiana Oliveira, and Amadeu M.V.M. Soares. Using a multibiomarker approach and behavioural responses to assess the effects of anthracene in *Palaemon serratus*. *Aquatic Toxicology*, 149:94–102, apr 2014.
- [3] Terminal de São Sebastião completa 40 anos. http://www.transpetro.com.br/pt{_}br/imprensa/noticias/terminal-de-sao-sebastiao-completa-40-anos.html. Accessed:2017-11-07.
- [4] Ronaldo J. Torres, Augusto Cesar, Camilo D. S. Pereira, Rodrigo B. Choueri, Denis M. S. Abessa, Marcos R. L. do Nascimento, Pedro S. Fadini, and Antonio A. Mozeto. Bioaccumulation of Polycyclic Aromatic Hydrocarbons and Mercury in Oysters (*Crassostrea rhizophorae*) from Two Brazilian Estuarine Zones. *International Journal of Oceanography*, 2012:1–8, jul 2012.
- [5] United States Environmental Protection Agency USEPA. Priority Pollutant List. <https://www.epa.gov/eg/toxic-and-priority-pollutants-under-clean-water-act>, 2014. 2017-06-19.
- [6] L.R. Vieira, A. Sousa, M.F. Frasco, I. Lima, F. Morgado, and L. Guilhermino. Acute effects of Benzo[a]pyrene, anthracene and a fuel oil on biomarkers of the common goby *Pomatoschistus microps* (Teleostei, Gobiidae). *Science of The Total Environment*, 395(2-3):87–100, jun 2008.
- [7] Xiaoke Hu, Peng Wang, and Huey-min Hwang. Oxidation of anthracene by immobilized laccase from *Trametes versicolor*. *Bioresource Technology*, 100(21):4963–4968, nov 2009.
- [8] X.D. Huang, D.G. Dixon, and B.M. Greenberg. Increased Polycyclic Aromatic Hydrocarbon Toxicity Following Their Photomodification in Natural Sunlight: Impacts on the Duckweed *Lemna gibba* L. G-3. *Ecotoxicology and Environmental Safety*, 32(2):194–200, nov 1995.

- [9] J.T. Oris, J.P. Giesy, P.M. Allred, D.F. Grant, and P.F. Landrum. Photoinduced Toxicity of Anthracene in Aquatic Organisms: an Environmental Perspective. *Studies in Environmental Science*, 25:639–658, 1984.
- [10] Luis Alberto Sáenz, Edson Luiz Seibert, Julianio Zanette, Haidi Dalida Fiedler, Adilson José Curtius, Jaime Fernando Ferreira, Eduardo Alves de Almeida, Maria Ri-soleta Freire Marques, and Afonso Celso Dias Bainy. Biochemical biomarkers and metals in *Perna perna* mussels from mariculture zones of Santa Catarina, Brazil. *Ecotoxicology and Environmental Safety*, 73(5):796–804, jul 2010.
- [11] Walter Narchi and Mario Sérgio Galvão-Bueno. Anatomia funcional de *Perna perna* (Linné) (bivalvia, Mytilidae). *Revista Brasileira de Zoologia*, 14(1):135–168, 1997.
- [12] B.S. Pierri, T.D. Fossari, and A.R.M. Magalhães. O mexilhão *Perna perna* no Brasil: nativo ou exótico? *Arquivo Brasileiro de Medicina Veterinária e Zootecnia*, 68(2):404–414, apr 2016.
- [13] Lílían Nogueira, Danielly Garcia, Rafael Trevisan, Ana Letícia Madeira Sanches, Daiane da Silva Acosta, Alcir Luiz Dafre, Thiago Yukio Kikuchi Oliveira, and Eduardo Alves de Almeida. Biochemical responses in mussels *Perna perna* exposed to diesel B5. *Chemosphere*, 134:210–216, sep 2015.
- [14] Jeferson L. Franco, Daniela B.B. Trivella, Rafael Trevisan, Daniel F. Dinslaken, Maria R.F. Marques, Afonso C.D. Bainy, and Alcir L. Dafre. Antioxidant status and stress proteins in the gills of the brown mussel *Perna perna* exposed to zinc. *Chemico-Biological Interactions*, 160(3):232–240, apr 2006.
- [15] Priscilla TY Leung, Jack CH Ip, Sarah ST Mak, Jian Qiu, Paul KS Lam, Chris KC Wong, Leo L Chan, and Kenneth MY Leung. De novo transcriptome analysis of *Perna viridis* highlights tissue-specific patterns for environmental studies. *BMC Genomics*, 15(1):804, 2014.
- [16] Mackenzie R. Gavary and Steven B. Roberts. Characterizing short read sequencing for gene discovery and RNA-Seq analysis in *Crassostrea gigas*. *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics*, 7(2):94–99, jun 2012.
- [17] P. A. Zagatto and E. Bertolotti. *Ecotoxicologia Aquática: Princípios e Aplicações*. RIMA, 2006.
- [18] F. A Azevedo and A. A. M. Chasin. *As bases toxicológicas da ecotoxicologia*. RIMA, 2004.
- [19] Changjiang Xu, Christina Yong-Tao Li, and Ah-Ng Tony Kong. Induction of phase I, II and III drug metabolism/transport by xenobiotics. *Archives of pharmacal research*, 28(3):249–68, mar 2005.
- [20] B Ewing, L Hillier, M C Wendl, and P Green. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome research*, 8(3):175–85, mar 1998.
- [21] Zheng Chang, Guojun Li, Juntao Liu, Yu Zhang, Cody Ashby, Deli Liu, Carole L Cramer, and Xiuzhen Huang. Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. *Genome Biology*, 16(1):30, 2015.

- [22] Yuan Zhang, Yanni Sun, James R. Cole, DG Brown, and JD Neufeld. A Scalable and Accurate Targeted Gene Assembly Tool (SAT-Assembler) for Next-Generation Sequencing Data. *PLoS Computational Biology*, 10(8):e1003737, aug 2014.
- [23] M. H. Schulz, D. R. Zerbino, M. Vingron, and E. Birney. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28(8):1086–1092, apr 2012.
- [24] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W Birren, Chad Nusbaum, Kerstin Lindblad-Toh, Nir Friedman, and Aviv Regev. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotech*, 29(7):644–652, jul 2011.
- [25] Rob Patro, Stephen M Mount, and Carl Kingsford. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat Biotech*, 32(5):462–464, may 2014.
- [26] Raghu Chandramohan, Po-Yen Wu, John H. Phan, and May D. Wang. Benchmarking RNA-Seq quantification tools. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 647–650. IEEE, jul 2013.
- [27] Nicolas L Bray, Harold Pimentel, Pall Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nat Biotech*, 34(5):525–527, may 2016.
- [28] Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Meth*, 14(4):417–419, apr 2017.
- [29] Avi Srivastava, Hirak Sarkar, Nitish Gupta, and Rob Patro. RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes. *Bioinformatics*, 32(12):i192–i200, jun 2016.
- [30] Fatemeh Seyednasrollah, Asta Laiho, and Laura L Elo. Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in bioinformatics*, 16(1):59–70, jan 2015.
- [31] Yue Him Wong, Jin Sun, Li Sheng He, Lian Guo Chen, Jian-Wen Qiu, and Pei-Yuan Qian. High-throughput transcriptome sequencing of the cold seep mussel *Bathymodiolus platifrons*. *Scientific Reports*, 5(1):16597, dec 2015.
- [32] Rebeca Moreira, Patricia Pereiro, Carlos Canchaya, David Posada, Antonio Figueras, and Beatriz Novoa. RNA-Seq in *Mytilus galloprovincialis*: comparative transcriptomics and expression profiles among different tissues. *BMC Genomics*, 16(1):728, dec 2015.
- [33] Meiyong Xu, Lihua Jiang, Kang-Ning Shen, Changwen Wu, Guangyuan He, and Chung-Der Hsiao. Transcriptome response to copper heavy metal stress in hard-shelled mussel (*Mytilus coruscus*). *Genomics data*, 7:152–4, mar 2016.

- [34] Xiu Jiang, Liguu Qiu, Hongwei Zhao, Qinqin Song, Hailong Zhou, Qian Han, and Xiaoping Diao. Transcriptomic responses of *Perna viridis* embryo to Benzo(a)pyrene exposure elucidated by RNA sequencing. *Chemosphere*, 163:125–132, nov 2016.
- [35] Badreddine Sellami, Abdelhafidh Khazri, Amine Mezni, Hla Louati, Mohamed Delali, Patricia Aissa, Ezzeddine Mahmoudi, Hamouda Beyrem, and David Sheehan. Effect of permethrin, anthracene and mixture exposure on shell components, enzymatic activities and proteins status in the Mediterranean clam *Venerupis decussata*. *Aquatic Toxicology*, 158:22–32, jan 2015.
- [36] W D MacLeod, D W Brown, A J Friedman, D G Burrow, O Mayes, R W Pearce, C A Wigren, and R G Bogar. *Standard analytical procedures of the NOAA national analytical facility 1985–1986. Extractable toxic organic compounds*. Number October 1985. U. S. Department of Commerce, NOAA/NMFS, second edition, 1985.
- [37] K. Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, jul 2002.
- [38] Itamar Sela, Haim Ashkenazy, Kazutaka Katoh, Tal Pupko, Oh T.K., Schneider D., Lenski R.E., Kim J.F., Wilm A., and Lopez R. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Research*, 43(W1):W7–W14, jul 2015.
- [39] Alexandros Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)*, 30(9):1312–3, may 2014.
- [40] A Rambaut. FigTree. <http://tree.bio.ed.ac.uk/software/figtree/>, 2016. 2017-06-26.
- [41] Babraham Bioinformatics. FastQC A Quality Control tool for High Throughput Sequence Data. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. 2017-06-26.
- [42] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, aug 2014.
- [43] W. Li and A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, jul 2006.
- [44] Felipe A. Simo, Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212, oct 2015.
- [45] Brian J Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, Matthew D MacManes, Michael Ott, Joshua Orvis, Nathalie Pochet, Francesco Strozzi, Nathan Weeks, Rick Westerman, Thomas William, Colin N Dewey, Robert Henschel, Richard D LeDuc, Nir Friedman, and Aviv Regev. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protocols*, 8(8):1494–1512, aug 2013.

- [46] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat Meth*, 9(4):357–359, apr 2012.
- [47] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, oct 1990.
- [48] Rachid Ounit, Steve Wanamaker, Timothy J Close, and Stefano Lonardi. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*, 16(1):236, dec 2015.
- [49] Robert C. Edgar. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461, oct 2010.
- [50] P. Jones, D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A. F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S.-Y. Yong, R. Lopez, and S. Hunter. InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240, may 2014.
- [51] Robert D Finn, Jody Clements, and Sean R Eddy. HMMER web server: interactive sequence similarity searching. *Nucleic acids research*, 39(Web Server issue):W29–37, jul 2011.
- [52] Yuki Moriya, Masumi Itoh, Shujiro Okuda, Akiyasu C Yoshizawa, and Minoru Kanehisa. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic acids research*, 35(Web Server issue):W182–5, jul 2007.
- [53] Dan Tenenbaum. KEGGREST: Client-side REST access to KEGG. <https://bioconductor.org/packages/release/bioc/html/KEGGREST.html>, 2016. Accessed:2017-03-04.
- [54] Maria Murgarella, Daniela Puiu, Beatriz Novoa, Antonio Figueras, David Posada, Carlos Canchaya, and RK Aziz. A First Insight into the Genome of the Filter-Feeder Mussel *Mytilus galloprovincialis*. *PLOS ONE*, 11(3):e0151561, mar 2016.
- [55] Nicole G. Ertl, Wayne A. O’Connor, Alexie Papanicolaou, Aaron N. Wiegand, Abigail Elizur, and M Robles. Transcriptome Analysis of the Sydney Rock Oyster, *Saccostrea glomerata*: Insights into Molluscan Immunity. *PLOS ONE*, 11(6):e0156649, jun 2016.
- [56] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, may 2000.
- [57] Eduardo A. Almeida, Afonso Celso Dias Bainy, Alcir Luiz Dafre, Osmar F. Gomes, Marisa H.G. Medeiros, and Paolo Di Mascio. Oxidative stress in digestive gland and gill of the brown mussel (*Perna perna*) exposed to air and re-submersed. *Journal of Experimental Marine Biology and Ecology*, 318(1):21–30, may 2005.

- [58] Raymond. Gilles and Michelle. Gilles-Baillien. *Transport Processes, Iono- and Osmoregulation : Current Comparative Approaches*. Springer Berlin Heidelberg, 1985.
- [59] Regina Coimbra Rola, Marta Marques Souza, and Juliana Zomer Sandrini. Hypoosmotic stress in the mussel *Perna perna* (Linnaeus, 1758): Is ecological history a determinant for organismal responses? *Estuarine, Coastal and Shelf Science*, 189:216–223, apr 2017.
- [60] Guofan Zhang, Xiaodong Fang, Ximing Guo, Li Li, Ruibang Luo, Fei Xu, Pengcheng Yang, Linlin Zhang, Xiaotong Wang, Haigang Qi, Zhiqiang Xiong, Huayong Que, Yinlong Xie, Peter W H Holland, Jordi Paps, Yabing Zhu, Fucun Wu, Yuanxin Chen, Jiafeng Wang, Chunfang Peng, Jie Meng, Lan Yang, Jun Liu, Bo Wen, Na Zhang, Zhiyong Huang, Qihui Zhu, Yue Feng, Andrew Mount, Dennis Hedgecock, Zhe Xu, Yunjie Liu, Tomislav Domazet-Loso, Yishuai Du, Xiaoqing Sun, Shoudu Zhang, Binghang Liu, Peizhou Cheng, Xuanting Jiang, Juan Li, Dingding Fan, Wei Wang, Wenjing Fu, Tong Wang, Bo Wang, Jibiao Zhang, Zhiyu Peng, Yingxiang Li, Na Li, Jinpeng Wang, Maoshan Chen, Yan He, Fengji Tan, Xiaorui Song, Qiumei Zheng, Ronglian Huang, Hailong Yang, Xuedi Du, Li Chen, Mei Yang, Patrick M Gaffney, Shan Wang, Longhai Luo, Zhicai She, Yao Ming, Wen Huang, Shu Zhang, Baoyu Huang, Yong Zhang, Tao Qu, Peixiang Ni, Guoying Miao, Junyi Wang, Qiang Wang, Christian E W Steinberg, Haiyan Wang, Ning Li, Lumin Qian, Guojie Zhang, Yingrui Li, Huanming Yang, Xiao Liu, Jian Wang, Ye Yin, and Jun Wang. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature*, 490(7418):49–54, oct 2012.
- [61] Ximing Guo, Yan He, Linlin Zhang, Christophe Lelong, and Aude Jouaux. Immune and stress responses in oysters with insights on adaptation. *Fish and Shellfish Immunology*, 46(1):107–119, 2015.
- [62] Grace Chen, Michael H. Shaw, Yun-Gi Kim, and Gabriel Nuñez. NOD-Like Receptors: Role in Innate Immunity and Inflammatory Disease. *Annual Review of Pathology: Mechanisms of Disease*, 4(1):365–398, feb 2009.
- [63] KEGG PATHWAY Database. <http://www.genome.jp/kegg/pathway.html>, 2017. Accessed:2017-09-18.
- [64] Identification and function of a novel C1q domain-containing (C1qDC) protein in triangle-shell pearl mussel (*Hyriopsis cumingii*). *Fish & Shellfish Immunology*, 58:612–621, nov 2016.
- [65] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, jan 2010.
- [66] Sonya M. Billiard, Mark E. Hahn, Diana G. Franks, Richard E. Peterson, Niels C. Bols, and Peter V. Hodson. Binding of polycyclic aromatic hydrocarbons (PAHs) to teleost aryl hydrocarbon receptors (AHRs). *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, 133(1):55–68, sep 2002.
- [67] Ryeo-Eun Go, Kyung-A Hwang, and Kyung-Chul Choi. Cytochrome P450 1 family and cancers. *The Journal of Steroid Biochemistry and Molecular Biology*, 147:24–30, mar 2015.

- [68] Natalija Marinković, Daria Pasalić, and Slavica Potocki. Polymorphisms of genes involved in polycyclic aromatic hydrocarbons' biotransformation and atherosclerosis. *Biochemia medica*, 23(3):255–65, 2013.
- [69] Negar Gharavi and Ayman O.S. El-Kadi. Down-regulation of aryl hydrocarbon receptor-regulated genes by tumor necrosis factor- α and lipopolysaccharide in murine hepatoma hepa 1c1c7 cells. *Journal of Pharmaceutical Sciences*, 94(3):493–506, mar 2005.
- [70] Hesham M. Korashy and Ayman O.S. El-Kadi. The role of aryl hydrocarbon receptor and the reactive oxygen species in the modulation of glutathione transferase by heavy metals in murine hepatoma cell lines. *Chemico-Biological Interactions*, 162(3):237–248, sep 2006.
- [71] Kim F. Rewitz, Bjarne Styris have, Anders Løbner-Olesen, and Ole Andersen. Marine invertebrate cytochrome P450: Emerging insights from vertebrate and insect analogies. *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology*, 143(4):363–381, aug 2006.
- [72] William S Baldwin, Peter B Marko, and David R Nelson. The cytochrome P450 (CYP) gene superfamily in *Daphnia pulex*. *BMC Genomics*, 10(1):169, apr 2009.
- [73] I. Boutet, A. Tanguy, and D Moraga. Characterisation and expression of four mRNA sequences encoding glutathione S-transferases pi, mu, omega and sigma classes in the Pacific oyster *Crassostrea gigas* exposed to hydrocarbons and pesticides. *Marine Biology*, 146(1):53–64, dec 2004.
- [74] D Sheehan, G Meade, V M Foley, and C A Dowd. Structure, function and evolution of glutathione transferases: implications for classification of non-mammalian members of an ancient enzyme superfamily. *Biochemical Journal*, 360(Pt 1):1–16, nov 2001.
- [75] Bae-Hoon Kim, Avinash R Shenoy, Pradeep Kumar, Clinton J Bradfield, and John D MacMicking. IFN-inducible GTPases in host cell defense. *Cell host & microbe*, 12(4):432–44, oct 2012.
- [76] Yanan Tian, Arnold B Rabson, and Michael A Gallo. Ah receptor and NF- κ B interactions: mechanisms and physiological implications. *Chemico-Biological Interactions*, 141(1-2):97–115, sep 2002.
- [77] Jennifer L Head and B Paige Lawrence. The aryl hydrocarbon receptor is a modulator of anti-viral immunity. *Biochemical pharmacology*, 77(4):642–53, feb 2009.
- [78] Edward T. Morgan. Regulation of Cytochrome P450 by Inflammatory Mediators: Why and How? *Drug Metabolism and Disposition*, 29(3), 2001.

Anexo A

Abaixo se encontra os relatórios gerados pelo software FastQC, que analisou a qualidade obtida no sequenciamento para cada amostra.

Figura A1: Análise da qualidade do sequenciamento por base da *read* 1, realizado pelo FastQC da amostra 1 do grupo exposto.

Figura A2: Análise da qualidade do sequenciamento por base da *read* 2, realizado pelo FastQC da amostra 1 do grupo exposto.

Figura A3: Análise da qualidade do sequenciamento por base da *read* 1, realizado pelo FastQC da amostra 2 do grupo exposto.

Figura A4: Análise da qualidade do sequenciamento por base da *read* 2, realizado pelo FastQC da amostra 2 do grupo exposto.

Figura A5: Análise da qualidade do sequenciamento por base da *read* 1, realizado pelo FastQC da amostra 3 do grupo exposto.

Figura A6: Análise da qualidade do sequenciamento por base da *read* 2, realizado pelo FastQC da amostra 3 do grupo exposto.

Figura A7: Análise da qualidade do sequenciamento por base da *read* 1, realizado pelo FastQC da amostra 4 do grupo exposto.

Figura A8: Análise da qualidade do sequenciamento por base da *read* 2, realizado pelo FastQC da amostra 4 do grupo exposto.

Figura A9: Análise da qualidade do sequenciamento por base da *read* 1, realizado pelo FastQC da amostra 1 do grupo controle.

Figura A10: Análise da qualidade do sequenciamento por base da *read* 2, realizado pelo FastQC da amostra 1 do grupo controle.

Figura A11: Análise da qualidade do sequenciamento por base da *read* 1, realizado pelo FastQC da amostra 2 do grupo controle.

Figura A12: Análise da qualidade do sequenciamento por base da *read* 2, realizado pelo FastQC da amostra 2 do grupo controle.

Figura A13: Análise da qualidade do sequenciamento por base da *read* 1, realizado pelo FastQC da amostra 3 do grupo controle.

Figura A14: Análise da qualidade do sequenciamento por base da *read* 2, realizado pelo FastQC da amostra 3 do grupo controle.

Figura A15: Análise da qualidade do sequenciamento por base da *read* 1, realizado pelo FastQC da amostra 4 do grupo controle.

Figura A16: Análise da qualidade do sequenciamento por base da *read* 2, realizado pelo FastQC da amostra 4 do grupo controle.

Anexo B

Abaixo se encontra os 20 termos de maior abundância no transcrito da brânquia do mexilhão marrom do *Perna perna* em cada ontologia do GO, que não está presente nos capítulos anteriores.

Figura B1: Os 20 termos GO mais abundantes na categoria BP a nível de nucleotídeo. O eixo X representa o número de transcritos associados

Figura B2: Os 20 termos GO mais abundantes na categoria MF a nível de nucleotídeo. O eixo X representa o número de transcritos associados

Figura B3: Os 20 termos GO mais abundantes na categoria CC a nível de nucleotídeo. O eixo X representa o número de transcritos associados

Figura B4: Os 20 termos GO mais abundantes na categoria CC a nível de proteína. O eixo X representa o número de transcritos associados