MINISTERIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL RURAL DA AMAZONIA-UFRA
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA APLICADA A AGROPECUÁRIA-PPGBAA

ROSYELY DA SILVA OLIVEIRA

MONTAGEM, PREDIÇÃO GÊNICA E ANOTAÇÃO DO GENOMA DA *Cynoscion acoupa* (Lacepède, 1801)

BELÉM-PA
2024

ROSYELY DA SILVA OLIVEIRA

MONTAGEM, PREDIÇÃO GÊNICA E ANOTAÇÃO DO GENOMA DA *Cynoscion acoupa* (Lacepède, 1801)

Dissertação apresentada ao Curso de Pós-Graduação em Biotecnologia Aplicada à Agropecuária, da Universidade Federal Rural da Amazônia como requisito para obtenção do título de Mestre.

Orientador: Dr. Marcus de Barros Braga
Coorientadora: Drª. Marília Danyelle Nunes Rodrigues

BELÉM-PA
2024

Dados Internacionais de Catalogação na Publicação (CIP)
Bibliotecas da Universidade Federal Rural da Amazônia
Gerada automaticamente mediante os dados fornecidos pelo(a) autor(a)

O48m Oliveira, Rosyely da Silva
MONTAGEM, PREDIÇÃO GÊNICA E ANOTAÇÃO DO GENOMA DA Cynoscion acoupa (Lacepède, 1801) / Rosyely da Silva Oliveira. - 2024.
65 f. : il. color.

Dissertação (Mestrado) - Programa de PÓS-GRADUAÇÃO em Biotecnolgia Aplicada à Agropecuária (PPGBAA), Campus Universitário de Belém, Universidade Federal Rural Da Amazônia, Belém, 2024.
Orientador: Prof. Dr. Marcus de Barros Braga
Coorientador: Profa. Dra. Marília Danyelle Nunes Rodrigues .

1. Bioinformática. 2. Genoma. 3. montagem. 4. pescado . I. Braga, Marcus de Barros, *orient.* II. Título

---

CDD 660.6

ROSYELY DA SILVA OLIVEIRA

MONTAGEM, PREDICAO GENICA E ANOTACAO DO GENOMA DA *Cynoscion acoupa* (Lacepede, 1801)

Dissertação apresentada à Universidade Federal Rural da Amazônia, como parte das exigências do Curso de Mestrado em Biotecnologia Aplicada à Agropecuária, para obtenção do título de Mestre. Área de concentração: Biotecnologia Animal.

Data da aprovação: 28/06/2024

Banca Examinadora

_____Orientador

Prof. Dr. Marcus de Barros Braga (UFRA- Paragominas)

_____Membro

Interno Prof. Dr. Fabricio Almeida Araújo (UFRA)

_____Membro

Interno Prof. Dra. Jakelyne Machado Lima Silva (UFRA)

_____Membro

Externo Prof. Dra. Regianne Maciel dos Santos Correa (UNIESAMAZ)

_____Membro

Suplente Prof. Dr. Rommel Thiago Juca Ramos – (UFPA)

Ao homem que sempre acreditou em mim e me deu forças, sempre foi meu guia, meu sol.

Te amo pai

AGRADECIMENTOS

Primeiramente gostaria de agradecer a Deus, segundamente ao CNPq por apoiar minha Dissertação e a Universidade Federal Rural da Amazônia por me dá essa grande oportunidade, juntamente com meu grande orientador Marcus Braga, obrigada por toda paciência, oportunidades, conhecimento e tempo que o senhor dedicou, espero um dia ser uma professora pelo menos parecida com você.

Agradeço ao meu pai, por que ele foi o grande "culpado" por eu chegar aqui, sempre me apoiou de todas as maneiras imagináveis e inimagináveis para que eu sempre priorizasse os estudos, me dando assim essa vontade de crescer e alcançar onde eu nem ao menos chegava a sonhar, e hoje sei graças ao senhor que o céu é o limite.

Você não poderia ficar de fora, Ricardo, obrigada por ser meu companheiro e sempre está ao meu lado no meio de todos os surtos e caos que minha vida se tornou nesses anos, sempre me escutando e aconselhando, com toda a paciência do mundo ouvindo eu falar sobre um assunto que você não fazia ideia da existência, e por mais que no final insistisse em dizer que meu trabalho era sobre peixe frito, você sempre esteve lá nos momentos mais difíceis, sem nunca reclamar e me dando o apoio que precisei.

Claudia, Diego e Rony, vocês são minha família, sabem disso, sempre torceram por mim e pela minha vitória, nunca de forma egoísta, sempre cuidaram de mim e me escutaram quando estava no meu limite, até quando a conversar estava difícil ou eu estava em um ponto que nem consigo descrever, obrigada por todos os "estou aqui, pode falar", por todas as risadas, calunias, fofocas, jogos... por todos os momentos que não posso escrever aqui, mas vocês sabem bem que momentos são esses, sem dúvidas eu nunca conseguiria essa vitória sem vocês me ajudando a deixar isso mais leve.

Claudia, Diego, Rony e Ricardo, em vocês encontrei a família que sempre quis ter, meus grandes agradecimentos.

Com amor Rosy

RESUMO

A aquicultura no Brasil teve resultados positivos em 2023, onde a produção nacional alcançou 887.03 toneladas. O *Cynoscion acoupa* é um peixe popularmente conhecido como “pescada-amarela”, encontrado principalmente no litoral norte e tem um grande valor econômico principalmente para fins gastronômicos. Mais recentemente foi observado um material chamado "*isinglass*", encontrado na bexiga natatória da pescada amarela, que vem sendo utilizado como matéria prima em diversas áreas do mercado, como por exemplo na indústria farmacêutica, civil, alimentícia, entre outros. O trabalho tem como objetivo montar o genoma, fazer a predição de genes e realizar a anotação funcional da espécie *C. acoupa,* vale ressaltar que esses dados gnômicos ainda não foram realizados por ninguém. Realizando a coleta dos espécimes nas regiões de Salinópolis-PA, Bragança-PA e na costa do Amapá, posteriormente foi feita a coleta do DNA de 5 *C. acoupa*, onde foi escolhida a melhor amostra para dar prosseguimento ao estudo. Através dessas amostras, foi feito o sequenciamento do genoma completo dos indivíduos capturados, com duas bibliotecas *paired-end short insert* (2 x 250 pb) DNAseq construído com o Kit Illumina DNA prep, utilizando a plataforma de sequenciamento NovaSeq SP 6000 Illumina, produzindo cerca de 1,302 Gb de sequências de dados brutos. Posteriormente, foi utilizado o FASTQC para verificar a qualidade do sequenciamento. A partir das sequências geradas será feita a montagem *de novo* do genoma utilizando o montador ABYSS. O alinhador BOWTIE-2 será usado para alinhar os *contigs* e *scaffolds* com o genoma de uma espécie filogeneticamente próxima, a ferramenta GeneMark para a predição gênica e o software BUSCO para a anotação do genoma, para que enfim possa deposita-lo no NCBI. Através dos resultados, espera-se trazer novas informações sobre o sequenciamento do genoma e a sua qualidade, juntamente com a publicação desses dados para estudos futuros.

Palavras-Chave: montagem de genoma; Pescada-amarela; predição gênica.

ABSTRACT

Aquaculture in Brazil had positive results in 2023, where national production reached 887,03 tons. The *Cynoscion acoupa* is a fish popularly known as “pescada-amarela”, found mainly on the north coast and has a great economic value mainly for gastronomic purposes. More recently, a material called "isinglass" was observed, found in the swim bladder of this fish, which has been used as raw material in several areas of the market, such as the pharmaceutical, civil, food, among others. The purpose of this work is to perform the genome assembling, gene prediction and carry out functional annotation of the *C. acoupa* organism. DNA samples were extracted from 5 specimens in the regions of Salinópolis-PA, Bragança-PA and on the coast of Amapá. The best sample was chosen to continue the study. Through these samples, the complete genome of the captured individuals was sequenced, with two paired libraries paired-end short insert (2 x 250 bp) DNAseq built with the Illumina DNA prep Kit, using the NovaSeq SP 6000 Illumina sequencing platform, producing raw data. Subsequently, FASTQC were used to verify the quality of the sequencing. From the generated sequences the de novo genome assembly will be made using the ABYSS assembler. BOWTIE-2 will be used to align the contigs and scaffolds with the genome of a phylogenetically close species, GeneMark will be used for gene prediction and BUSCO for genome annotation, so that it can finally be deposited at the NCBI. Through the results, it is expected to bring new information about genome sequencing and its quality, together with the publication of these data for future studies.

Key Words: gene prediction; genome assembly; Pescada-amarela.

SUMÁRIO

1. INTRODUÇÃO

A espécie *Cynoscion acoupa* (Lacepède, 1801), também conhecido popularmente como pescada-amarela, é um peixe marinho-estuarino e habita em águas tropicais e subtropicais na região costeira da América do Sul (Guimarães, 2018). A *C. acoupa* é nectônica, ou seja, é um organismo que se movimenta constantemente, frequenta águas rasas e salobras dos estuários, porém também entra em rios de água doce (Guimarães, 2018). A *Cynoscion acoupa* pode ser capturada durante todo o ano, porém, a produção se intensifica nos meses de setembro e agosto, no período de seca (Ferreira *et al.,* 2020). Seu nome popular pode variar em alguns países. Na Venezuela a pescada-amarela é conhecida como "curvina", já na Guiana Francesa é chamada de "acoupa rouge" (Oliveira, *et al* 2020). Os estudos a respeito da *C. acoupa* ainda são escassos, com poucas informações genéticas disponíveis, onde nenhum gênero de sua espécie tem o seu genoma sequenciado. Ele possui pouco investimento, mesmo considerando o valor econômico envolvido (Ferreira *et al.,* 2020).

O estado do Pará possui cerca de 562 km de região costeira que abrange cerca de 123 comunidades, destacando regiões de igarapés, rios, manguezais e estuários, favorecendo assim a produção pesqueira e exportando 27,5% da produção pesqueira do Brasil (Oliveira, 2021).

Em certos sistemas de produção pesqueira, são encontrados subprodutos valiosos, como a bexiga natatória de peixes, que é exportada a preços elevados, contribuindo para o aumento da receita na cadeia de produção de pesca (Medeiros, 2019). A *C. acoupa* é uma espécie que apresenta pele, órgãos e escamas com uma alta concentração de colágeno, (Monte, 2017). A partir dessa matéria-prima, é possível realizar a extração do *isinglass,* conhecido em alguns lugares no Brasil como ictiocola (colágeno extraído da bexiga natatória), que tem ampla aplicação na indústria (Costa et al., 2018).

A Bioinformática vem crescendo de forma exponencial a partir da década de 90 e está relacionada com a análise computacional de sequências de DNA, RNA e proteínas, unindo a biologia molecular com sistemas de informação (Santos, 2022). A biotecnologia e a bioinformática desempenham um papel vital e crescente na ciência e na sociedade moderna. A biotecnologia envolve o uso de organismos vivos, seus sistemas ou processos para desenvolver tecnologias inovadoras em diversas áreas (Oyawoye *et al*., 2022). A bioinformática, por sua

vez, combina biologia com ciência da computação, permitindo a análise, interpretação e armazenamento de grandes conjuntos de dados genéticos e moleculares, acelerando a pesquisa cientifica (Ferreira, 2018).

A composição do genoma e a anotação funcional são dois passos fundamentais no estudo de genomas, envolvendo organismos procariotos ou eucariotos (Charllis *et al.,* 2020). Ambos desempenham papéis fundamentais na compreensão da estrutura e da função dos genes e na realização de pesquisas na biologia molecular e biologia genômica (Vaser *et al.,* 2021). A montagem do genoma envolve a criação de uma sequência contínua e organizada dos nucleotídeos que compõem o DNA de um organismo (Giani *et al.,* 2020). Isso é essencial porque o DNA de muitos organismos, incluindo humanos, é composto por bilhões de pares de bases e está organizado em múltiplos cromossomos (Charllis *et al.,* 2020). A montagem do genoma permite: compreender a estrutura genômica; comparação entre espécies; identificação de variações genéticas (Giani *et al.,* 2020).

A anotação funcional é o processo de atribuir funções para genes e elementos não codificantes identificados no genoma (Giuffra *et al.,* 2019). Isso inclui a identificação de sequências codificadoras de proteínas, elementos regulatórios, RNAs não codificantes e outras características importantes (Li *et al.,* 2020). A anotação funcional é importante por entender as funções dos genes; identifica elementos regulatórios; compreende processos biológicos; facilita a pesquisa (Giuffra *et al.,* 2019).

A predição gênica é um passo importante para a anotação de novas sequências e genomas montados, porque através da predição gênica é possível analisar e buscar sequências de nucleotídeos que correspondem a cada um de seus genes ou de outras regiões de interesse (Lin *et al.,* 2020; Junior, 2018).

Com o avanço da tecnologia na ciência, viu-se a necessidade de estudar mais a fundo, em nível genômico, diversos organismos já conhecidos, como o pescado, que possui um grande valor econômico e gastronômico (De Meira *et al.,* 2017). Ao explorar as complexidades genéticas desse precioso recurso marinho, abre-se um vasto leque de possibilidades para aprimoradas não apenas a indústria pesqueira, mas também a compreensão das funções genéticas de outros pescados (Madeira, 2008). Comparar as características genômicas entre espécies distintas nos permitirão entender melhor a evolução e a adaptação desses seres vivos, capacitando-nos a utilizar de forma mais inteligente e sustentável os recursos oferecidos pelo vasto ecossistema aquático.

2. OBJETIVOS

2.1. Objetivo Geral

Realizar a montagem *de novo*, a predição de genes e a anotação funcional da espécie *Cynoscion acoupa*.

2.2. Objetivos Específicos

- Sequenciar o DNA do pescado *Cynoscion acoupa*;
- Realizar a montagem *de novo* do genoma do pescado *Cynoscion acoupa*;
- Efetuar a predição de genes no genoma da *Cynoscion acoupa;*
- Efetuar a anotação funcional do pescado *Cynoscion acoupa*;

3. REVISÃO DE LITERATURA

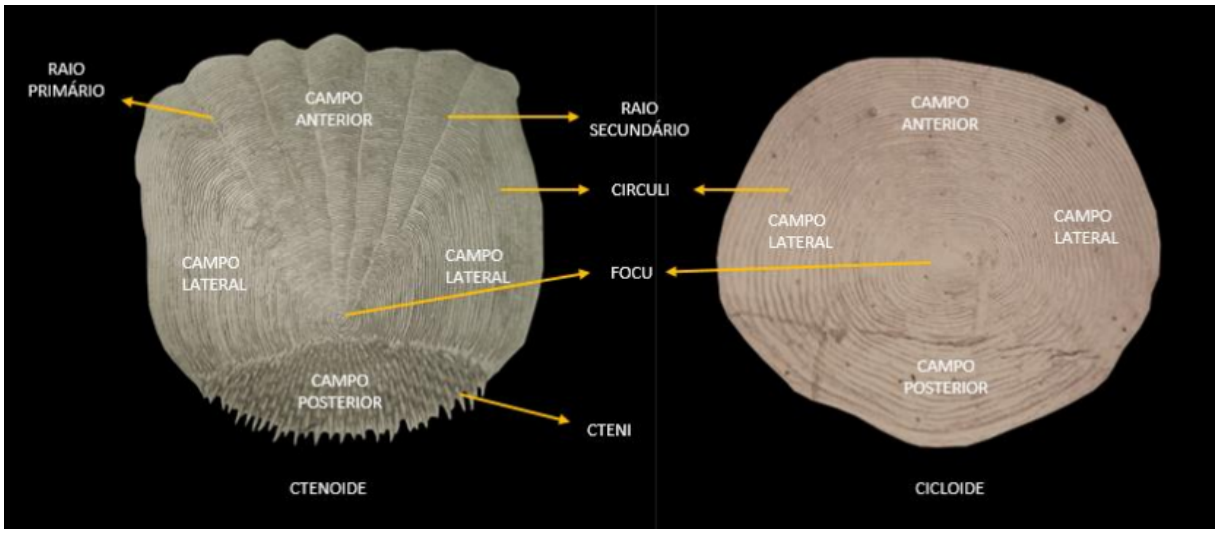
Por definição, pescado é qualquer organismo aquático que pode ser consumido, com isso, a pescada-amarela se enquadra nessa categoria (Barbosa, 2016). A *Cynoscion acoupa* é o maior representante da família *Sciaenidae*, ainda se tratando de uma espécie comercial importante para a América do Sul, sendo responsável por uma grande movimentação no mercado pesqueiro (Ferreira, *et al*., 2016).

A espécie *Cynoscion acoupa* (Lacepède, 1801) (Figura 1), pertencente ao filo Chordata, classe Actinopterygii, gênero *Cynoscion*, da Ordem Gadiformes, possui em sua morfologia a nadadeira dorsal com 11 espinhos e com até 22 raios; nadadeira anal com 2 espinhos e com até 9 raios; nadadeiras pélvicas, caldais e peitorais de coloração amarelada, e seu queixo não possui barbilhões ou esporos, possui escamas ctenóides, que estão presentes em sua maioria nos peixes ósseos. Estes peixes possuem uma textura áspera no corpo e escamas cicloides na cabeça, cujo formato é mais liso e circular, atualmente, percebe-se um aumento de interesse nos estudos de suas escamas que são uma das fontes de colágeno (figura 2) (Lemos, 2017). Atualmente existe poucos dados genômicos disponibilizados sobre o organismo de estudo em questão, seu genoma ainda não foi sequenciado, e suas regiões mais pesquisadas são proteínas e nucleotídeos.

Figura 1. Imagem da espécie *Cynoscion acoupa.*

Fonte: Junior (2017, p.20).

Figura 2. Imagem demonstrativa da morfologia dos 2 tipos diferentes de escamas que a *Cynoscion acoupa* possui.



Fonte: Mendes (2019).

O pescado já faz parte da culinária, sendo consumido semanalmente pela maioria da população brasileira, desde 2004, o consumo de pescado no Brasil cresceu 65%, em 2023, o Brasil importou US$ 1,4 bilhão em pescado (Granda, 2023). Comparado ao século passado, quando havia um consumo de cerca de 8,5 milhões de toneladas anualmente, houve um grande aumento no consumo do pescado no século XIX para cerca de 160 milhões de toneladas/ano (Barbosa, 2016).

3.1 Importância Econômica e Distribuição Geográfica

A bexiga natatória dos peixes é um órgão localizado em seu abdômen que possui a função de facilitar sua natação regulando a profundidade (Oliveira; Oliveira, 2020). A bexiga natatória é conhecida popularmente como “bucho” ou “grude”, e sua comercialização vem crescendo de forma exponencial (Medeiros, 2019). A *C. acoupa* possui pele, órgãos e escamas ricos em colágeno e diversos estudos apontam que é possível fazer extração desse material (Monte, 2017). A partir dessa matéria prima, é possível fazer a extração do *isinglass* (colágeno extraído da bexiga natatória) que pode ser amplamente utilizado na indústria (Costa *et al*., 2018).

O valor da bexiga natatória pode variar dependendo de país, do estado da matéria-prima e para qual seja o objetivo final. Ela passa por diversos procedimentos industriais até se obter uma “gelatina de alta qualidade”, que serve de matéria prima para a indústria na clarificação de cervejas e vinhos, na produção de cápsulas na indústria farmacêutica, na produção de textura em sucos e gelatinas, e até mesmo no preparo de cimento na construção civil. Vale ressaltar que

o colágeno extraído destes peixes também é utilizado na indústria de cosméticos e na fotografia, agregando assim um alto valor comercial para esse subproduto (Pinheiro, 2021).

A *Cynoscion acoupa,* é uma espécie de peixe que é de grande importância econômica em várias regiões, especialmente na América do Sul e Central (De matos; Lucena, 2011). O grude proveniente da pescada-amarela vem demonstrando seu alto valor econômico desde a década de 90, aumentando em 35% a pesca do *C. acoupa,* contudo, embora o grude tenha alto valor por quilograma em comparação à carne, é esta última que impulsionou majoritariamente os ganhos, representando 65% do rendimento econômico (Mourão, 2009).

Em algumas regiões a pescada-amarela é exportada para o mercado internacional, o que gera matéria prima para os países produtores. Essas exportações podem contribuir significativamente para o movimento comercial de um país. A disponibilidade de pescada-amarela e de outros pescados em áreas costeiras atrai entusiastas da pesca esportiva. O turismo de pesca pode gerar receita adicional por meio de pacotes turísticos, aluguel de barcos, guias de pesca e instrutores (De Alma *et al.,* 2011).

A carne da pescada-amarela é valorizada por sua qualidade e sabor suave. É uma fonte de proteína magra e é frequentemente consumida fresca ou congelada. É um peixe popular nos mercados de frutos do mar, restaurantes e supermercados (Mourão *et al.,* 2009). A pescada-amarela é uma fonte de alimento importante para muitas pessoas em regiões costeiras. Ela desempenha um papel na segurança alimentar, fornecendo uma fonte confiável de proteína (Santos, 2011).

A carne da pescada-amarela é versátil na culinária. Pode ser preparada de várias maneiras, como grelhada, assada, cozida, frita ou ensopada (Moura *et al.,* 2023). Sua natureza suave permite que ela absorva sabores e temperos, tornando-a uma excelente base para uma variedade de pratos (Ayulo; Machado; Scussel, 1994). Devido às suas características sensoriais e nutricionais, a carne da pescada-amarela é amplamente comercializada nos mercados de frutos do mar em várias partes do mundo (Moura *et al.,* 2023). A carne da pescada-amarela é uma excelente fonte de proteína magra, o que a torna uma escolha saudável para aqueles que procuram uma dieta equilibrada. Além disso, ela é rica em nutrientes essenciais, como selênio, fósforo, vitaminas do complexo B (como B12) e ômega-3, que são benéficos para a saúde (Santos, 2011).

A demanda por pescada-amarela cria oportunidades econômicas para pescadores, processadores de frutos do mar, distribuidores e restaurantes. Isso contribui para a economia local e regional em áreas onde a pesca desse peixe é significativa (Rodrigues *et al.,* 2020). Como fonte confiável de alimento, a pescada-amarela desempenha um papel importante na segurança

alimentar, fornecendo uma fonte de proteína acessível e nutritiva para comunidades costeiras e consumidores em geral (Freire *et al.,* 2019).

No mundo todo as principais fontes de extração do colágeno são de origem suína e bovina, porém, nas últimas décadas, os peixes vêm ganhando foco nessa área (Quintero; Zapata, 2017). No pescado é possível se obter 3 tipos de proteínas diferentes, entre elas temos a sarcoplasmática, estroma e miofibrilar (Dangaram *et al.,* 2009). A proteína estroma é responsável pelo colágeno e elastina (Zavarete, 2012), por se tratar de uma proteína capacitada em se envolver em diversos processos industriais, valorizando o subproduto, trazendo assim um foco mais recente a área de estudo (Prestes, 2013).

Pode-se observar a importância da pescada amarela em diversas áreas diferentes da indústria, bem como o seu valor comercial, e não apenas o pescado em si, mas também do seu genoma que ainda é pouco estudado. Mesmo com tamanha importância e envolvimento nas matérias-primas, o desenvolvimento de pesquisas nessa área é escarço, em especial, estudos aprofundados sobre os genes do colágeno do pescado através de ferramentas da bioinformática (Costa *et al.,* 2018).

3.2 Genômica

O genoma é o conjunto genético de um ser vivo, o número total de genes presentes nos cromossomos de um organismo, contendo informação genética (Kenehisa, 2023). Esse termo foi criado em 1920 por Hans Winkler, docente da área de botânica na Universidade de Hamburgo, Alemanha. A genômica é o campo de estudo que tem concentração no sequenciamento, mapeamento e análise do genoma, com o objetivo de compreender a organização, função e estrutura dos genes e da informação genética (Urbano; Braga, 2016) (Montelione; Anderson, 1999).

A estrutura genômica se dedica ao estudo da organização e estrutura do genoma, incluindo quadros de leitura aberta, RNA de transferência e RNA ribossômico (Kenehisa, 2023). A identificação da localização de um gene e outros marcadores auxilia na compreensão da função dos elementos genômicos (Terwilliger *et al*., 1998). Essa identificação é representada por mapas genéticos definidos por marcadores moleculares. Esses mapas são gerados com base em frequências de recombinação gênica, distância física molecular ou características citológicas dos cromossomos (Urbano; Braga, 2016).

Conhecer a sequência de DNA é uma etapa importante para entender como o genoma funciona. Por isso, a genômica funcional possibilita a análise da função dos genes em diferentes contextos, como desenvolvimento ou em resposta a mudanças ambientais. Além disso, existem outras áreas derivadas dessas subdivisões da genômica (De Keersmaecker *et al.,* 2006).

A Bioinformática foi pioneira em unificar a biologia com ciências computacionais, visando estudar nucleotídeos juntamente com suas funções e estrutura, principalmente com o desenvolvimento das tecnologias, o que ocasionaram os sequenciamentos de nova geração (NGS – Next-Generation Sequencing) (Verli, 2014). Com a chegada de uma nova era, os NGS ajudaram a avançar na investigação de micro e macro organismos, facilitando e deixando a bioinformática mais acessível para diferentes pesquisadores em todo o mundo (Boers e*t al.,* 2019). Através das plataformas de sequenciamento de nova geração é possível fazer uma grande leitura de dados em menos tempo e com mais eficácia (Mosele *et al.,* 2020). Com o avanço dos NGS também é possível auxiliar pesquisas mais massivas utilizando outros métodos biológicos, como por exemplo o PCR (Reação em Cadeia da Polimerase) (Tan *et al*., 2015).

A área da Bioinformática tem experimentado um rápido crescimento a partir dos anos 2000, que foi quando surgiu os primeiros sequenciadores de nova geração, e ela está intimamente ligada à análise computacional de sequências de DNA, RNA e proteínas, integrando os campos da biologia molecular e da tecnologia da informação, criando assim uma gama enorme de diversos programas capazes de fazer leituras de genes e bancos de dados de armazenamento de sequências biológicas (Santos, 2022). Com isso, tornou-se possível compreender de forma mais completa os dados biológicos gerados e expressos por essas ferramentas (Ferreira, 2018).

3.2.1 Montagem de Genomas

A espécie *Cynoscion acoupa* até o momento não teve seu genoma sequenciado e montado. Ela pertence ao gênero *Cynoscion*, onde dentro dele temos o total de 24 espécies filogeneticamente próximas, e nenhuma delas também teve seu genoma montado e disponibilizado em bancos de dados públicos. Já a sua família Sciaenidae possui 65 gêneros, e dentre eles 12 possuem espécies com genomas sequenciados e publicados em bancos de dados biológicos (National Center for Biotechnology information, 2023).

O sequenciamento de DNA começou com os métodos mais antigos, como o sequenciamento de Sanger, que permitiam a leitura de fragmentos relativamente curtos de DNA

de cada vez (Klasberg *et al.,* 2019). No entanto, avanços tecnológicos recentes, como a próxima geração de sequenciamento (NGS) e o sequenciamento de terceira geração, como a tecnologia de nanoporos, revolucionaram o campo da bioinformática (Jeon *et al.,* 2021). A NGS é uma abordagem de alto rendimento que permite a sequenciação simultânea de milhões de fragmentos de DNA em uma execução única (Klasberg *et al.,* 2019). Isso resulta na geração de enormes quantidades de dados sequenciais, que precisam ser processados, analisados e interpretados usando ferramentas de Bioinformática (Behjati; Tarpey, 2013).

A tecnologia de sequenciamento tem uma ampla gama de aplicações na bioinformática. Ela é utilizada para sequenciar o genoma completo de organismos, incluindo humanos, plantas, animais e micro-organismos, o que fornece informações valiosas sobre a estrutura genética e as variações entre os indivíduos (Hu *et al*., 2021). Além disso, o sequenciamento é usado para estudar o transcriptoma, que é o conjunto de lista de RNA expressos em uma célula ou tecido específico, permitindo a análise da expressão gênica e a descoberta de novos genes (Hu *et al*., 2021). Também é utilizado para identificar doenças genéticas, como câncer, Alzheimer e rastrear a origem de doenças infecciosas, investigar a diversidade genética de indivíduos e realizar estudos evolutivos (Rodrigues, 2009).

A tecnologia de sequenciamento tem o potencial de revolucionar a medicina personalizada, permitindo a identificação de variantes genéticas associadas a doenças específicas e facilitando a seleção de tratamentos personalizados com base no perfil genético individual (Behjati; Tarpey, 2013). No entanto, o sequenciamento em si é apenas o primeiro passo. A análise e interpretação dos dados sequenciais requer o uso de ferramentas bioinformáticas sofisticadas, algoritmos e métodos estatísticos para realizar tarefas como montagem de genoma, anotação genômica, identificação de variantes, análise de expressão gênica e análise filogenética (Urbano; Braga, 2016).

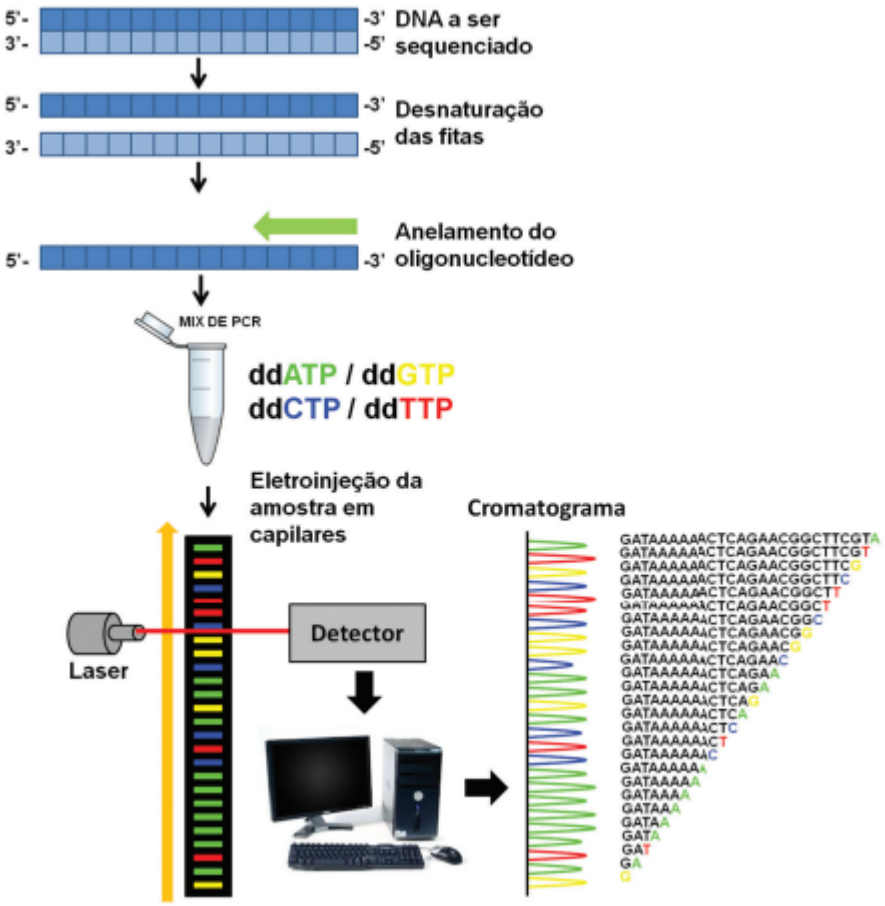
Contudo, a tecnologia de sequenciamento na Bioinformática vem desempenhando um papel crucial na geração de dados genômicos e transcriptômicos, permitindo a compreensão mais aprofundada dos processos biológicos, a descoberta de novos conhecimentos e aplicações em várias áreas, como medicina, agricultura, biologia evolutiva e biotecnologia (Jorge, 2016).

Dentre os diversos fabricantes de plataformas de sequenciamento NGS, a Illumina atualmente é uma das principais existente no mercado, com uma vasta gama de sequenciadores que podem ser encontrados, como o iSeq 100, MiSeq**,** HiSeq 2500, entre outros (Jeon *et al.,* 2021). Este tipo de sequenciamento NGS pode ser usado para determinar a composição bacteriana de comunidades microbianas complexas com o intuito de descobrimento de morfologia e funcionamento (Heikema *et al.,* 2020). A Illumina introduziu o HiSeq X Ten, com

células de fluxo padronizadas. A Illumina NGS aproveita o sequenciamento por síntese (SBS) procurando a introdução de nucleotídeos marcados à proporção que a cadeia de DNA é copiada (paralela) (Jeon *et al.,* 2021). Os NGS geram um número maior de dados de sequenciamento de DNA, de forma mais barata do que o sequenciamento tradicional que é o método Sanger (Hu *et al*., 2021). O sequenciamento Illumina pode fornecer uma saída de dados variando de 300 kilobases (kb) até vários terabases (Tb) em uma única execução (Stoler; Nekrutenko, 2021).

O processo de montagem de sequências genômicas tem como objetivo reconstruir o genoma alvo através da combinação de pequenas sequências chamadas *reads* e levando em consideração as sobreposições entre elas (El-Metwally *et al.,* 2013). Na década de 70, um novo método de sequenciamento foi descoberto, o método Sanger (figura 3), através dele é possível incorporar nucleotídeos alterados quimicamente, possibilitando assim posteriormente novas tecnologias como *Next Generation Sequencing* (NGS).

Figura 3 :Automatização do método de Sanger. Após a desnaturação da dupla fita, didesoxinucleotídeos marcados com compostos florescentes são incorporados à cadeia nascente de DNA sintetizada pela DNA polimerase.



Fonte: Moreira (2015).

Existem duas abordagens principais para realizar essa montagem: a montagem por referência, que utiliza um genoma de referência próximo filogeneticamente ao genoma alvo, e uma abordagem *de novo*, que realiza a montagem apenas com base nas sobreposições das leituras, sem a necessidade de uma referência (Urbano; Braga, 2016). Antes de iniciar a montagem, é necessário realizar o pré-processamento dos dados para lidar com possíveis erros de sequenciamento (Vrancken *et al.,*2016). Uma etapa importante nesse processo é a avaliação da qualidade das leituras, qualidade essa que é medida pela confiabilidade dos dados de sequências de nucleotídeos obtidos durante o processo de sequenciamento (De Sena, 2019). Essa avaliação permite identificar regiões de baixa qualidade por meio de histogramas, auxiliando na tomada de decisões quanto à remoção de bases de baixa qualidade (Wojcieszek *et al.,* 2014) (Urbano; Braga, 2016).

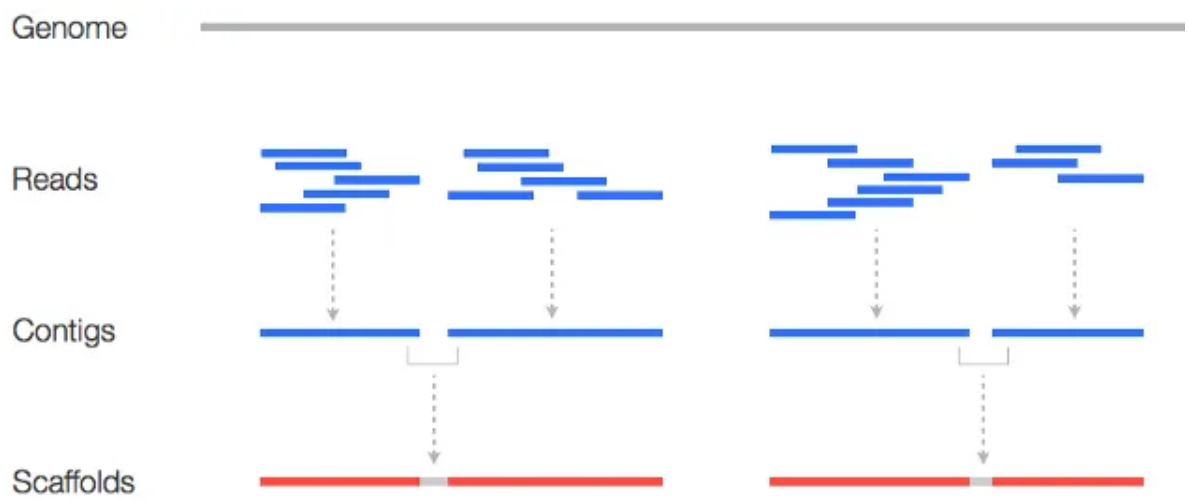
Com o pré-processamento, a montagem pode ser realizada por meio de diferentes abordagens, como algoritmos gulosos, consenso de layout de sobreposição (OLC, *overlap-layout-consensus*) e grafos de *De Bruijn* (Nagarajan; Pop, 2013). No algoritmo guloso, cada leitura é comparada com todas as outras em busca das melhores sobreposições, repetindo-se o processo até que todas as leituras sejam utilizadas (Nagarajan; Pop, 2013). Em seguida, as sobreposições são utilizadas para agrupar as leituras e formar sequências contíguas, também conhecidas como *contigs* (Urbano; Braga, 2016). Na abordagem OLC, o algoritmo também busca pela melhor sobreposição, porém, essa informação é organizada em um grafo no qual cada vértice representa uma leitura e as arestas que os conectam são as sobreposições entre essas leituras (Wojcieszek *et al..* 2014). Nos montadores que utilizam grafos *De Bruijn*, as leituras são divididas em sub-leituras de tamanho fixo, chamadas *k-mers* (Ramos, 2022). Estes *k-mers* são usados para gerar um grafo onde cada vértice representa um *k-mer* e as arestas são as sobreposições de tamanho k-1(Ramos, 2022). Nestes algoritmos, a etapa de busca pela melhor sobreposição não existe, o que resulta numa redução do esforço computacional (Nagarajan; Pop, 2013).

A montagem por referência é um método utilizado na bioinformática para reconstruir sequências de DNA a partir de leituras obtidas por sequenciamento. Este método baseia-se na comparação das leituras sequenciadas com uma sequência de referência previamente conhecida, como um genoma já sequenciado de uma espécie próxima ou da mesma espécie. A montagem por referência é amplamente utilizada em estudos genômicos devido à sua eficiência e precisão.

A etapa final da montagem consiste na geração dos *scaffolds* (Figura 4), que utiliza as leituras em pares para orientar e ordenar os *contigs* gerados durante o processo de montagem (Scott *et al.*, 2020). As distâncias estimadas entre as leituras pareadas ajudam a determinar a distância entre os *contigs* e a preencher as lacunas presentes nos *scaffolds* (Bayat *et al.*, 2018; Urbano; Braga, 2016). Alguns montadores possuem módulos automatizados para a geração dos *scaffolds* durante o processo de montagem (Scott *et al.,* 2020). Ao final da montagem, geralmente, são gerados dois arquivos de saída: um contendo os *contigs* e outro contendo os *scaffolds* (Bayat *et al.*, 2018; Urbano; Braga, 2016).

É possível que ainda ocorram lacunas após a geração do *scaffolds*, que podem ser diminuídas ou até mesmo totalmente resolvidas por programas específicos (Zhou *et al.,* 2023). Por fim, é feita a avaliação do resultado da montagem, etapa de extrema importância para o bom processamento pelos montadores (Zhou *et al.,* 2023). Programas de bioinformática podem avaliar diferentes métricas para diferentes montagens, sejam por referência ou *De Novo*, disponibilizam tabelas e gráficos com resultados comparativos, podem auxiliar nesta etapa (Mikheenko *et al*., 2018). Tecnologia de sequenciamento na bioinformática refere-se a um conjunto de técnicas e métodos utilizados para determinar a ordem precisa dos nucleotídeos em uma molécula de DNA ou RNA (Verli, 2014). Essa informação sequencial é fundamental para compreender a estrutura, a função e a variação genética dos organismos (Verli, 2014).

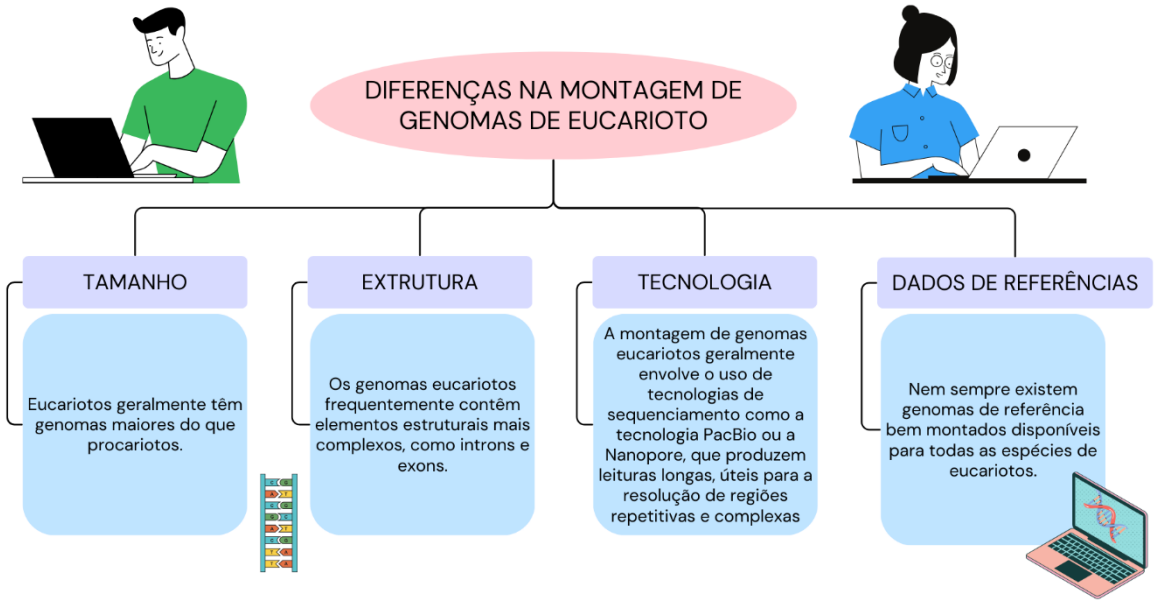
Figura 4: Processo de montagem de genomas: *Reads*, *Contigs* e *Scaffolds*.



Fonte: Kremer, (2020).

A montagem de genomas de eucariotos, em comparação com a montagem de genomas de procariotos, apresenta adicionais devido à maior complexidade e estruturação genética dos genomas eucarióticos (Tørresen *et al.,* 2019). Algumas das principais diferenças na montagem de genomas de eucariotos, com ênfase nas características específicas desses genomas, podem ser observadas na ilustração abaixo (Figura 5):

Figura 5**.** Principais diferenças da montagem de genomas eucariotos.



Fonte: Os autores (2024).

Em resumo, a montagem de genomas de eucariotos é mais desafiadora devido à complexidade estrutural, tamanho variável dos genomas e a presença de elementos repetitivos (Tørresen *et al.,* 2019). Isso requer o uso de técnicas avançadas de sequenciamento e montagem, bem como uma compreensão profunda da biologia molecular para montar com precisão os genomas eucarióticos e identificar genes funcionais (Li *et al.,* 2019; Tørresen *et al.,* 2019)

A montagem por referência é um método que utiliza um genoma de referência previamente conhecido ou uma sequência relacionada filogeneticamente como guia para a montagem das sequências alvo (Pop; Salzberg, 2008). Esse método é particularmente útil quando se tem um genoma de referência bem caracterizado disponível, como é o caso de organismos com genomas amplamente conhecidos, como humanos, moscas de frutas e bactérias comuns (Heller; Vingron, 2020). A montagem por referência envolve o controlado

das sequências curtas ou "reads" seguindo do sequenciamento contra o genoma de referência, procurando correspondências e sobreposições (Pop; Salzberg, 2008). Com base nesse acompanhamento, as leituras são organizadas e verificadas em sequências mais longas e contíguas, conhecidas como "*contigs*" (Heller; Vingron, 2020).

Por outro lado, a montagem *de novo* é um método que não requer um genoma de referência e realiza a montagem das sequências alvo apenas com base nas informações de sobreposição entre as leituras (Wee *et al*., 2019). Nesse método, como se lê são detectados entre si para identificar sobreposições e regiões de similaridade. Essas sobreposições são usadas para construir *contigs* maiores, que representam sequências genômicas contíguas (Wick *et al*., 2015). A montagem *de novo* é particularmente útil quando não há um genoma de referência adequado disponível ou quando o objetivo é obter uma sequência genômica completamente nova, como no caso de organismo pouco estudado ou em projetos de descoberta de novas espécies (Paszkiewicz; Studholme, 2010). No entanto, uma montagem *de novo* é mais desafiadora e computacionalmente intensiva, uma vez que requer uma resolução de sobreposições complexas e identificação de regiões repetitivas (Wee *et al.,* 2018).

Ambos os métodos têm suas limitações. A montagem por referência pode apresentar dificuldades em regiões genômicas altamente variáveis ou em genomas altamente divergentes em relação ao genoma de referência (Peker *et al.,* 2019). Além disso, pode não capturar variações genômicas ou inserções/deleções em relação ao genoma de referência (Peker *et al.,* 2019). Por outro lado, uma montagem *de novo* pode levar à formação de *contigs* incompletos ou contendo erros devido a sobreposições ambíguas ou repetições (Wee *et al.,* 2018; Heller; Vingron, 2020).

Os montadores de genoma são software e algoritmos utilizados para montar sequências de DNA a partir de dados brutos de sequenciamento (Khan et al., 2018). Essas ferramentas desempenham um papel crucial na genômica, permitindo a reconstrução de genomas completos de organismos, desde bactérias até seres humanos (Silva; Notaria; Dall'alba, 2020). O campo da montagem de genomas está em constante evolução, com novos algoritmos e ferramentas sendo desenvolvidos regularmente. Pesquisadores devem acompanhar as atualizações para garantir que estejam usando a melhor ferramenta disponível para suas necessidades (Khan *et al.,* 2018).

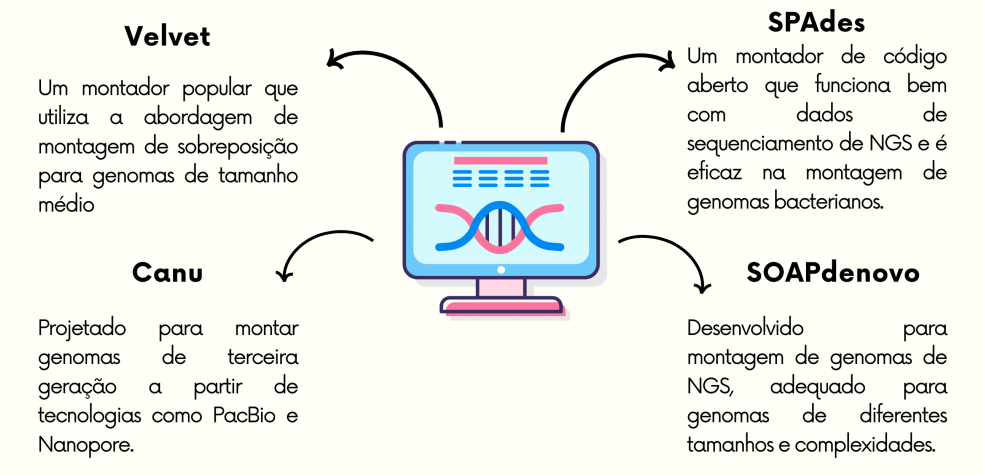
A montagem de genoma consiste na reconstrução das sequências de DNA, RNA e proteína a partir de fragmentos de sequenciamento, que podem ser obtidos por técnicas como o sequenciamento de nova geração (NGS) ou até mesmo o sequenciamento de terceira geração com a tecnologia PacBio ou Nanopore (Lu; Giordano; Ning, 2016). A montagem de genoma

pode ser desafiadora devido a fatores como repetições no genoma, variações de tamanho de sequência, erros de sequenciamento e a qualidade dos dados, que são fatores cruciais para uma montagem de qualidade (Alkan; Sajjadian; Eichler, 2011). Além disso, a complexidade do genoma (procariótico ou eucariótico) também influencia os desafios enfrentados na montagem (Alkan; Sajjadian; Eichler, 2011). A escolha do montador de genoma e a configuração de seus parâmetros são importantes para obter resultados precisos. Isso envolve ajustar os parâmetros de acordo com a qualidade dos dados de sequenciamento e a natureza do genoma (Sohn et al., 2018).

A montagem *de novo* ABySS funciona com dados experimentais fazendo análises de transcriptomas (Nikolic *et al.,* 2022). ABySS é um montador de sequência para leituras curtas, montando leituras de 36-50 pb (Birol *et al.,* 2009). Outra ferramenta muito utilizada na bioinformática é o Bowtie 2. Se trata de uma ferramenta utilizada para alinhar as leituras de sequenciamento com cerca de 50 até 100 ou 1000 pares de bases, também são bons em alinhamentos longos, como o dos mamíferos (Yao *et al.,* 2020). O Bowtie 2 também fornece suporte aos modos de alinhamentos que possuem lacunas, locais e extremidades emparelhadas (Sang, 2021).

Existem diversos montadores de genoma disponíveis, cada um com suas próprias características e vantagens. Alguns exemplos no esquema montado a baixo (figura 6):

Figura 6. Exemplos de montadores de genoma.



Fonte: Os autores (2024).

O SOAPdenovo é um software de montagem de genomas desenvolvido pelo BGI (Instituto Genômico de Pequim, Beijing Genomics Institute), uma das maiores organizações de sequenciamento de genomas existente. Este montador de genomas é projetado para a montagem

de sequências de DNA de alto rendimento geradas por tecnologias de sequenciamento de nova geração (NGS) (Li *et al.,* 2015).

O SOAPdenovo utiliza uma abordagem de montagem de sequência de sobreposição de leituras curtas, que é comumente empregada em montadores de genomas de NGS (Carpenter *et al.,* 2019). Ele analisa as sobreposições entre leituras curtas para montar contigs que, posteriormente, são usados para montar contigs maiores e, finalmente, o genoma completo (MADRITSCH *et al.,* 2021). Logo ele é escalável e pode ser usado para montar genomas de diferentes tamanhos, desde genomas de bactérias até genomas eucarióticos mais complexos, incluindo genomas de plantas e animais (Carpenter *et al.,* 2019).

O software oferece configurações personalizáveis, permitindo que os pesquisadores ajustem parâmetros, como tamanho mínimo de sobreposição, para se adequarem aos dados de sequenciamento específicos e aos genomas que estão sendo montados (Qin *et al*., 2021). O SOAPdenovo tem sido amplamente utilizado em pesquisas genômicas e é conhecido por seu bom desempenho na montagem de genomas, desde genomas bacterianos até genomas de organismos mais complexos (Kooij; Pellicer, 2020). No entanto, seu desempenho pode depender da qualidade dos dados de sequenciamento e da configuração adequada dos parâmetros (Zheng *et al.,*2019).

Além disso, o desenvolvimento de ferramentas de montagem de genomas está em constante evolução, desenvolvimento e novas versões ou alternativas podem estar disponíveis para atender às necessidades específicas dos pesquisadores (Khew *et al.,* 2020). Portanto, espera-se que ocorra a consulta a literatura científica e as fontes atualizadas para obter informações sobre as opções mais recentes em montagens de genomas (Miao *et al.,* 2021).

A montagem de genomas é fundamental para a pesquisa nas áreas da biologia, genética, biotecnologia e medicina. Ela é usada para estudos de variações genômicas, genomas de organismos não-modelo, diagnóstico de doenças genéticas e muito mais (Sandberg *et al.,* 2019). Em resumo, os montadores de genoma desempenham um papel fundamental na análise de sequências genômicas, permitindo a reconstrução e o estudo de genomas completos (Li *et al.,* 2023). A escolha do montador e a configuração adequada são cruciais para obter resultados precisos e úteis na pesquisa em que estiver trabalhando (De Jesus-Pires *et al.,* 2020).

Com o surgimento de toda essa nova tecnologia de sequenciamento, viu-se a necessidade de bancos de dados para armazenar todas as novas informações e que também possua acesso fácil e simples, como por exemplo o GenomeARK. Os princípios do GenomeArk é ser um espaço de trabalhos e repositório de banco de dados para genomas de referência de alta qualidade de todas as espécies (Howe, 2020). Os conjuntos finais depositados são curados

por especialistas antes de serem enviados aos bancos de dados públicos, como por exemplo o NCBI GenBank (Tuner *et al., 2023*). O GenomeArk pode ser acessado através do link (https://genomeark.github.io/) (Howe, 2020).

Porém, é necessário ter um controle sobre a qualidade dos sequenciamentos e montagens disponíveis nos bancos de dados. O BUSCO (Benchmarking Universal Single-Copy Orthologs) é uma ferramenta que tem como objetivo avaliar a montagem do genoma, se baseando no conceito de genes ortólogos de cópia única, onde são conservados entre as espécies relacionadas (Simão *et al.,* 2015). Como por exemplo, pesquisadores que procuram estudar a completude do genoma de mamífero deverão usar genes ortólogos de cópia única que já foram descobertos em outros mamíferos (Simão *et al.,* 2015).

3.2.2 Predição Gênica

A predição de genes é um passo importante para a anotação de novas sequências e genomas montados, e é responsável por analisá-las e buscar sequências de nucleotídeos correspondentes a cada um de seus genes ou de outras regiões de interesse (Mathé *et al.* 2002). O seu objetivo é encontrar a localização de diferentes partes que compõe a estrutura do gene.

A predição gênica é o processo de identificar regiões codificantes de proteínas (genes) dentro de uma sequência de DNA. Este processo é essencial para a anotação de novos genomas e envolve uma combinação de técnicas computacionais e experimentais. As abordagens principais para predição gênica incluem a predição baseada em similaridade e a predição *ab initio*. (Wang *et al*., 2004).

Na predição baseada em similaridade, utiliza-se sequências de proteínas ou genes conhecidos de outras espécies para encontrar regiões semelhantes na sequência de DNA em estudo. Ferramentas como o BLAST (Basic Local Alignment Search Tool) são amplamente utilizadas para realizar estas buscas de similaridade, ou o Exonerate que é um programa que executa alinhamentos de sequência de DNA e proteínas para prever a localização e estrutura de genes. Quando uma sequência semelhante é encontrada, a anotação do gene conhecido é transferida para a nova sequência. Esta abordagem é particularmente eficaz quando há genomas de referência bem anotados de espécies próximas (Rogalska *et al.,* 2023).

A predição *ab initio* utiliza características estatísticas e modelos matemáticos para identificar genes com base apenas na sequência de DNA. Algoritmos *ab initio* analisam a sequência de DNA em busca de padrões que correspondem a características conhecidas de

genes, como sinais de iniciação e terminação de transcrição, regiões promotoras, codões de início e término, e padrões de exons e íntrons (Stanke *et al.,* 2006). Ferramentas como GENSCAN, AUGUSTUS, e GeneMark são exemplos de programas *ab initio* que utilizam modelos probabilísticos, como modelos de Markov ocultos, para prever genes (Burge, Karlin, 1997; Stanke *et al*., 2004; Lomsadze et al., 2005).

A predição de genes moderna frequentemente integra múltiplas fontes de dados, incluindo RNA-Seq, dados de proteínas e informações epigenéticas. A combinação de métodos *ab initio* e baseados em similaridade, juntamente com dados experimentais, pode aumentar significativamente a precisão das predições (Trapnell *et al.,* 2010).

A predição gênica enfrenta vários desafios, como a presença de genes sobrepostos, onde uma sequência de DNA pode codificar mais de um gene, e a existência de splicing alternativo (é o processo pelo qual os íntrons são removidos do pré-mRNA e os éxons são unidos para formar um mRNA maduro que será traduzido em uma proteína, e quando ele é alternativo permite que diferentes combinações de éxons sejam juntadas, resultando em múltiplas variantes de mRNA a partir de um único gene) , onde diferentes exons são combinados para formar múltiplas variantes de mRNA (Black, 2000). Genomas complexos, com muitas regiões repetitivas, pseudogenes ou regiões altamente conservadas, também complicam a predição. Além disso, a precisão das ferramentas computacionais varia, e a integração com dados experimentais é frequentemente necessária para garantir previsões confiáveis (Lin *et al.,* 2020; Junior, 2018).

A predição gênica tem diversas aplicações, incluindo a anotação de novos genomas, a descoberta de novos genes associados a doenças e a compreensão de mecanismos genéticos. Na agricultura e biotecnologia, a predição gênica ajuda na identificação de genes de interesse para o melhoramento genético de plantas e animais (Zhou; Troyanskaya, 2015). Se trata de uma área dinâmica e em constante evolução, essencial para a biologia computacional e a genômica. Ferramentas e métodos modernos continuam a melhorar, integrando novas tecnologias e fontes de dados para fornecer previsões mais precisas e detalhadas. À medida que mais genomas são sequenciados, a importância de métodos precisos e eficientes de predição de genes continuará a crescer, impulsionando avanços em biotecnologia, medicina e biologia evolutiva.

3.2.3 Anotação Funcional

A anotação do genoma é o processo de identificar e atribuir funções a elementos genômicos, como genes, regiões regulatórias e outros elementos funcionais em um organismo (Reed et al., 2006). Isso envolve a identificação de sequências de DNA que codificam proteínas (genes), bem como a determinação de onde esses genes começam e terminam, quais éxons e íntrons eles contêm e qual é a sequência de aminoácidos da proteína que codificam (Reed et al., 2006).

A anotação estrutural descreve a localização precisa dos diferentes elementos em um genoma, como quadros de leitura abertos, regiões de codificação, éxons, íntrons, repetições, locais de splicing, sequências regulatórias, códons de início e parada e promotores (Oliveira, 2019). A anotação funcional atribui funções aos elementos genômicos encontrados pela anotação estrutural, relacionando-os a processos biológicos como o ciclo celular, morte celular, desenvolvimento, metabolismo, etc. identificando elementos que possam ter sido anotados por erro (Burgarelli *et al.,* 2018).

Além da identificação de genes, a anotação do genoma também pode envolver a identificação de sequências regulatórias, como promotores, que controlam a expressão gênica (Seemann, 2014). Também pode incluir a identificação de elementos repetitivos, regiões não codificantes do genoma e características estruturais, como telômeros e centrômeros (Yandell *et al.,* 2012).

A anotação do genoma é essencial para entender a função e a estrutura dos genomas dos organismos, bem como para realizar estudos comparativos entre diferentes espécies (Seemann, 2014). Auxiliando os pesquisadores a elucidar os mecanismos moleculares subjacentes a processos biológicos e a desenvolver aplicações em áreas como biotecnologia, medicina e conservação. Atualmente no mercado existe algumas ferramentas de bioinformática que nos possibilitam a fazer a anotação (Tatusova *et al.,* 2016).

A anotação funcional busca identificar a função biológica da sequência do genoma e também pode levar a descobertas em organismos de interesse, assim como também é possível alcançar descobertas relacionadas a estrutura e funcionalidade de organismos através da anotação funcional. Pode-se compreender a anotação funcional como um processo que auxilia diversas descobertas de componentes que possuem suma importância nos organismos de estudos, principalmente genes e seus produtos (Santos, 2022).

A anotação do genoma é um processo computacional que envolve uma associação de informações relevantes para a biologia aos dados genômicos sequenciados (Santos, 2022). Recentemente, houve um avanço significativo na automação desse processo. A anotação compreende a identificação e descrição de genes, proteínas, vias regulatórias e metabólicas (Shumate; Salzberg, 2021). Geralmente, isso é realizado por meio de *pipelines* de anotação, que são combinação de softwares usados em sequência. Às vezes, o processo envolve o elemento humano para lidar com as anotações geradas automaticamente, o que é conhecido como curadoria manual (Médigue; Moszer, 2007).

O processo de anotação pode ser dividido em duas fases. A primeira fase, chamada de fase computacional ou automática, utiliza várias fontes de evidência de genomas ou dados transcriptômicos de espécies relacionadas para realizar uma predição inicial de genes e transcrições. Na fase subsequente, conhecida como curadoria manual, todas as informações anotadas durante uma fase automática são revisadas e resumidas em uma anotação final. Geralmente, as informações são manipuladas de um genoma de referência próximo para a nova sequência usando métodos baseados em homologia (Richardson; Watson, 2012; Urbano; Braga, 2016).

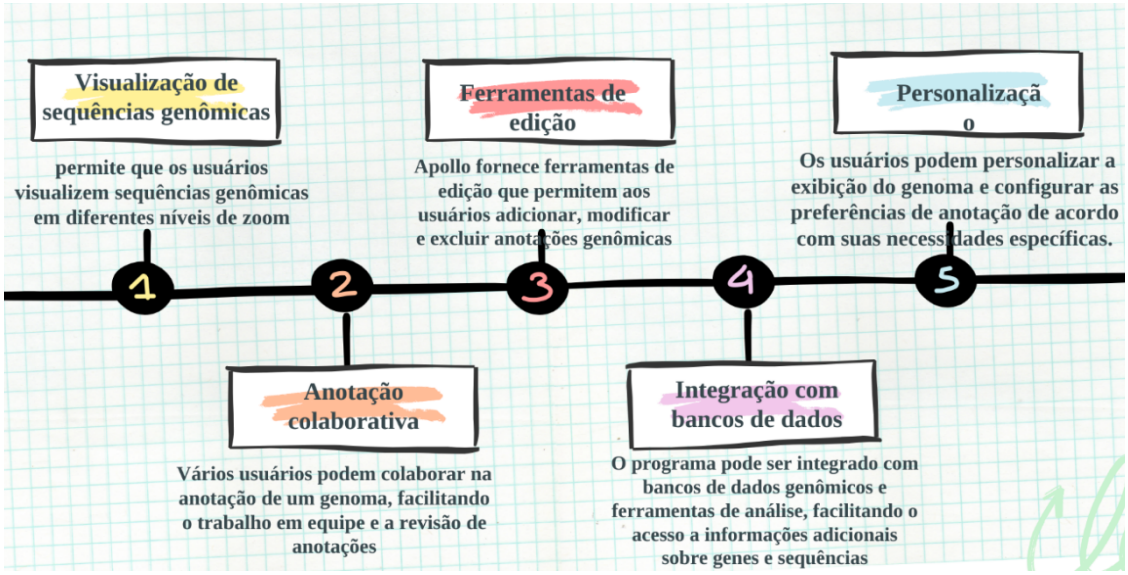
Com o aumento exponencial dos sequenciamentos de genomas, há menos tempo disponível para a anotação manual, o que leva a uma maior dependência de *pipelines* de anotação automática (Urbano; Braga, 2016). No entanto, o uso exclusivo de *pipelines* controlados pode resultar na introdução e controle de erros, como erros de ortografia, nomes idênticos a genes com produtos diferentes e diferenciados entre ortólogos e parálogos (Urbano; Braga, 2016). Esses erros podem levar a anotações inconsistentes e incorretas. Portanto, a curadoria manual desempenha um papel crucial na detecção e correção desses erros. Como resultado deste aumento no sequenciamento e montagem de genomas, houve um consequente aumento no número de genomas anotados disponíveis em bancos de dados públicos (Stothard; Wishart, 2006).

Uma das características distintivas do GeneMark é sua capacidade de levar em consideração a presença de estruturas de introns-exons em eucariotos, o que o torna particularmente útil na anotação de genomas de organismos mais complexos (Hoff *et al.,*2016). O GeneMark é frequentemente usado em projetos de anotação de genomas eucarióticos, tanto de organismos modelo quanto de espécies de interesse agronômico, médico ou ecológico (Mcininch *et al.,* 1996). Sua precisão, velocidade e adaptabilidade o tornam uma ferramenta valiosa na genômica comparativa e funcional (Hoff *et al.,*2016).

O GeneMark já foi responsável pela anotação funcional de diversos genes de eucariotos, entre eles temos a *Danio rerio*, também conhecido como zebra fish, que é um organismo modelo amplamente utilizado em pesquisas em biologia do desenvolvimento e genética (Arick *et al.,* 2023). Seu genoma foi anotado usando o GeneMark, entre outras ferramentas, para identificar genes e regiões regulatórias importantes (Arick *et al.,* 2023). O *Tetraodon nigroviridis,* este é um peixe amplamente estudado em biologia evolutiva, e o GeneMark foi empregado na anotação de seu genoma para identificar genes e elementos regulatórios (Brůna *et al.,*2020)

O Apollo é uma ferramenta de software de código aberto desenvolvida pelo Berkeley Bioinformatics Open-Source Projects (BBOP) para a anotação manual de sequências genômicas (Dun *et al.,* 2019). O principal objetivo do Apollo é permitir que os pesquisadores identifiquem e anotem genes e outras características genômicas em sequências de DNA (Lee *et al.,* 2009). Ele oferece uma interface gráfica amigável que permite aos usuários visualizarem o genoma, identificarem genes e marcadores, bem como adicionarem anotações e atributos a esses elementos (Firtina *et al.,* 2020). Com diversas características e funcionalidades (figura 7)

Figura 7. Principais características e funcionalidades do programa Apollo.



Fonte: Os autores (2024).

Em resumo, o Apollo é uma ferramenta poderosa e amplamente utilizada na anotação manual de sequências genômicas, desempenhando um papel importante na compreensão da estrutura e função dos genomas de diferentes organismos (Firtina *et al.,* 2020).

O BLAST (Basic Local Alignment Search Tool) é uma ferramenta fundamental em bioinformática, frequentemente aplicada na anotação de genomas (Oehmen; Nieplocha, 2006).

Embora o BLAST seja mais conhecido por sua função de alinhamento de sequências, ele desempenha um papel crucial na anotação de genomas, especialmente na identificação de genes e na atribuição de funções aos elementos genômicos (Syngai *et al.,* 2013).

Na anotação do genoma, o BLAST é usado para comparar sequências de DNA ou proteína recém-identificadas com sequências de referência em bancos de dados públicos, como o GenBank ou o UniProt. Isso permite aos pesquisadores inferir a função de genes recém-descobertos com base em sua similaridade com genes previamente caracterizados (Gupta *et al.,* 2014).

Por exemplo, após a predição de um novo gene por meio de ferramentas de predição de genes como o GeneMark ou o AUGUSTUS, os pesquisadores podem usar o BLAST para comparar a sequência de aminoácidos da proteína codificada por esse gene com sequências de proteínas conhecidas em bancos de dados públicos (Gabler *et al.,* 2020). Se houver uma alta similaridade com uma proteína previamente caracterizada, isso pode fornecer pistas sobre a função do novo gene. Além disso, o BLAST também pode ser usado para identificar regiões conservadas, elementos regulatórios e outras características funcionais nos genomas, contribuindo assim para uma anotação mais abrangente e detalhada (Edwards; Cottage, 2001). Em resumo, o BLAST é uma ferramenta essencial na anotação de genomas, ajudando os pesquisadores a atribuir funções aos genes recém-identificados e a compreender melhor a estrutura e a função dos genomas de uma variedade de organismos (Kaudal *et al.,* 2006).

O Gnomon é mais uma ferramenta de anotação de genes desenvolvida pelo National Center for Biotechnology Information (NCBI) que é frequentemente usada para identificar genes em genomas eucarióticos (Legeai *et al.,* 2010). Ele usa uma variedade de métodos, incluindo alinhamentos de sequências homólogas, modelagem de HMM (Hidden Markov Models) e algoritmos de predição de genes ab initio para identificar e anotar genes em sequências de DNA (Nagy *et al.,* 2008).

O Gnomon foi aplicado em uma ampla variedade de espécies de eucariotos para a anotação de seus genomas (Legeai *et al.,* 2010). No caso de espécies de peixes, embora o Gnomon possa ter sido usado em alguns estudos, é importante notar que a anotação do genoma geralmente envolve uma combinação de várias ferramentas e abordagens, incluindo o Gnomon, GeneMark, AUGUSTUS, EVidenceModeler (EVM), entre outros (Hoff *et al.,* 2016). Alguns exemplos de espécies de peixes cujos genomas foram anotados usando o Gnomon incluem: *Danio rerio, Tetraodon nigroviridis* e *Gasterosteus aculeatus*. Vale ressaltar que a combinação de diferentes ferramentas e abordagens é essencial para produzir uma anotação completa e precisa do genoma de uma espécie (Legeai *et al.,* 2010).

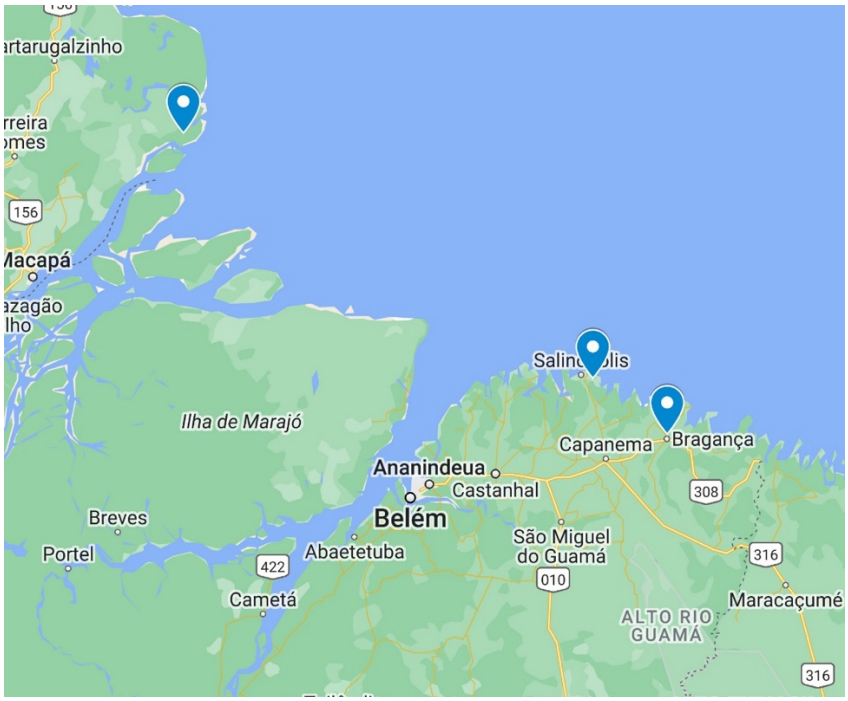
Com o avanço da tecnologia na ciência, viu-se a necessidade de estudar mais a fundo, em nível genômico, diversos organismos já conhecidos, como por exemplo o pescado, que possui um grande valor econômico e gastronômico, buscando assim ter uma maior compreensão sobre as funções dos genes que possui, e como melhor usá-los ao nosso favor.

4. METODOLOGIA

4.1 Coleta das Amostras e Extração de DNA

Os peixes utilizados neste trabalho foram capturados de três locais: Cuiarana, vila do município de Salinópolis/PA (Latitude: -0.631339, Longitude: -47.3461), Bragança/PA (Latitude: -1.06595, Longitude: -46.7895) e na costa do estado do Amapá (Latitude: 2.05108, Longitude: -50.7945) (Figura 8). Essas 3 regiões foram escolhidas por serem localidades de fácil acesso para os pescadores, por possuírem proximidade de portos e mercados de peixe, o que facilita a captura, transporte e comercialização dos peixes, e também essas áreas são conhecidas por terem uma alta concentração de Pescada-Amarela, devido ter presença de habitats adequados, como estuários e zonas costeiras ricas em nutrientes, favorece a reprodução e o crescimento desses peixes. As amostras de pescadas amarelas capturadas na costa do Amapá, foram coletadas no desembarque pesqueiro no mercado do Ver-o-Peso, em Belém/PA são exibidas na Figura 9. O DNA foi coletado do tecido muscular (nadadeira ventral) dos peixes. O material biológico coletado foi analisado através do procedimento de eletroforese em gel, pelo laboratório da Universidade Federal Rural da Amazônia (UFRA) – Belém, e a amostra de melhor qualidade foi escolhida para dar prosseguimento na metodologia de sequenciamento.

Figura 8. Localização da Origem das Amostras Coletadas.



Fonte: Google Mapas (2023).

Figura 9. Peixes de origem da costa do Amapá.

Fonte: Os autores (2023).

4.2 Sequenciamento do Genoma

Foi realizado o sequenciamento do genoma completo dos indivíduos capturados com duas bibliotecas pareadas *paired-end short insert* (2 x 250 pb) DNAseq construídas com o Kit Illumina DNA prep, utilizando a plataforma de sequenciamento NovaSeq SP 6000 Illumina, produzindo 1,302 Gb de sequências (dados brutos). A integridade das amostras de DNA coletadas e sequenciadas pode ser observada na Figura 11.

Posteriormente foi realizada a avaliação do sequenciamento utilizando o FastQC (Andrews *et al.,* 2004) para medir a qualidade das leituras geradas. A Tabela 1 descreve os parâmetros das bibliotecas pareadas.

Tabela 1. Bibliotecas *paired-end*.

ID	CODIFICAÇÃO	TOTAL DE SEQUÊNCIAS	TOTAL DE BASES	TAMANHO DAS SEQUÊNCIAS	%GC
PA-VP-3_S1_L001_R1_001	Sanger / Illumina 1.9	270.582.401	63.7 Gbp	35-251	42
PA-VP-3_S1_L001_R2_001	Sanger / Illumina 1.9	270.582.401	63.7 Gbp	35-251	42
PA-VP-3_S1_L002_R1_001	Sanger / Illumina 1.9	380.666.644	90.1 Gbp	35-251	42

PA-VP-3_S1_L002_R2_001	Sanger / Illumina 1.9	380.666.644	90.1 Gbp	35-251	42

4.3 Montagem do Genoma

Utilizou-se o software KmerGenie (Chikhi; Medvedev, 2014) com parâmetros padrão para estimar o melhor tamanho de k-mer (k) para a montagem. Foi escolhido o KmerGenie por nos entregar os melhores resultados e por se tratar de um software que possui bom desempenho de alta cobertura, que já vem sendo validado por outros trabalhos de genomas de eucariotos. O tamanho correto do k-mer é crucial durante o processo de montagem de genomas, pois afeta diretamente a precisão e a eficiência da reconstrução do genoma original a partir dos fragmentos de sequência. A escolha do tamanho correto de k-mer tem impacto direto na resolução de sobreposições, discriminação de regiões repetitivas, complexidade computacional e no tamanho dos contigs. Portanto, escolher o tamanho correto do k-mer é uma consideração crítica na montagem de genomas, pois pode afetar significativamente a qualidade e a completude do genoma reconstruído. Uma escolha cuidadosa do tamanho do k-mer pode ajudar a maximizar a precisão da montagem e a eficiência dos recursos computacionais utilizados.

A ferramenta apontou o tamanho 71 como melhor valor de k para a montagem das leituras da pescada amarela.

Com base nesse valor de k-mer, utilizou-se o montador MEGAHIT (Li *et al.,* 2016) (tabela 2) por apresentar alto desempenho em genomas grandes, algoritmo avançado e compatibilidade com várias plataformas. Para assim realizar a montagem *de novo* da pescada amarela com os seguintes parâmetros:

Tabela 2. Parâmetros utilizados no montador MEGAHIT.

Parâmetros	Valor
-1	/home/rommelufpa/pescada/PA-VP-3_S1_L001_R1_001. fastq,/home/rommelufpa/pescada/PA-VP-3_S1_L002_R1_001.fastq
-2	/home/rommelufpa/pescada/PA-VP 3_S1_L001_R2_001. fastq,/home/rommelufpa/pescada/PA-VP-3_S1_L002_R2_001.fastq
--k-list	71
-o	/tmp/megahit_pescada
--tmp-dir	/tmp/temp_megahit
--out-prefix	pescada_mega
-t	64
MEGAHIT_TEMP_DIR	/tmp/temp_megahit/megahit_tmp_EXNT1j/

Sendo:

- Arquivos de entrada (-1 e -2): Os caminhos dos arquivos *fastq* que contêm as sequências de DNA pareadas para a montagem do genoma. Os arquivos -1 e -2 correspondem aos pares de leitura 1 e leitura 2, respectivamente.
- Lista de k-meros (--k-list): Define a lista de valores de k-mer a serem testados durante o processo de montagem. No caso, o valor específico foi 71.
- Diretório temporário (--tmp-dir): O diretório onde serão armazenados os arquivos temporários durante o processo de montagem.
- Prefixo de saída (--out-prefix): Define o prefixo do nome dos arquivos de saída gerados pelo MEGAHIT. No caso, os arquivos de saída terão o prefixo "pescada_mega".
- Número de threads (-t): O número de threads (ou processadores) a serem utilizados durante o processo de montagem. Neste caso, foram utilizados 64 threads.

Esses parâmetros são essenciais para configurar e executar o processo de montagem do genoma utilizando o MEGAHIT.

A avaliação da montagem foi feita com a ferramenta QUAST (Gurevich *et al*., 2013). Já a estimativa de k-mer e a montagem foram realizadas em um computador com processador AMD Opteron (TM) Processor 6238 de 2600 MHz, 24 cores (utilizando 64 threads) e 2 TB de memória RAM.

4.4 Predição Gênica

Após a montagem do genoma, a predição dos genes foi feita com o software AUGUSTUS (Stanke *et al.,* 2006) em um computador com processador Intel(R) Xeon(R) Silver 4210R de 2.40 GHz, 30 cores e 72GB de memória RAM. Utilizou-se os parâmetros padrão da ferramenta, com exceção do parâmetro “species”; neste caso, utilizou-se o valor “zebrafish”. Foi escolhido esse parâmetro por se tratar da espécie de peixe mais utilizada na bioinformática, devido ao seu genoma bem caracterizado, transparência durante seu desenvolvimento, similaridade genética com humanos e sua aplicação em diversas áreas, pois foi utilizado em pesquisas que vão desde o desenvolvimento e regeneração até estudos de doenças genéticas e screening de drogas.

Existem diversos programas capazes de realizar a predição gênica, como o AUGUSTUS que é um programa de predição de genes *ab initio* amplamente utilizado em

bioinformática para anotação de genomas eucarióticos (Hoff *et al.,* 2016). Desenvolvido por Mario Stanke e Oliver Keller, o AUGUSTUS é conhecido por sua capacidade de prever genes de eucariotos de forma precisa e eficiente (Hoff *et al.,* 2019). O programa baseia-se em modelos estatísticos de Markov ocultos (HMMs) para identificar genes em sequências de DNA. Ele leva em consideração características como composição de bases, distribuição de codões e padrões de splicing para melhorar a precisão das previsões de genes (Brůna *et al.,* 2023).

Uma das principais vantagens do AUGUSTUS é sua capacidade de adaptação a diferentes espécies. Ele pode ser treinado com dados de sequenciamento genômico de uma espécie específica, o que melhora significativamente sua precisão na predição de genes para aquela espécie em particular (Stanke *et al.,* 2004). O AUGUSTUS é amplamente utilizado em projetos de anotação de genomas eucarióticos, contribuindo para a identificação de genes e para a compreensão da estrutura e função dos genomas de diversas espécies (Stanke *et al.,* 2003). Sua precisão, adaptabilidade e facilidade de uso fazem dele uma ferramenta valiosa na área da genômica comparativa e funcional. A espécie mais conhecida que fez uso do AUGUSTUS na anotação do genoma é o *Homo sapiens*, o genoma humano foi extensivamente anotado usando várias ferramentas de anotação, incluindo o AUGUSTUS, para identificar genes e regiões regulatórias importantes (Hoff *et al.,* 2019).

O GeneID é outra ferramenta de predição de genes amplamente utilizada na anotação de genomas eucarióticos. Desenvolvido por Jaime E. Blair e Pablo Librado, o GeneID é projetado para identificar genes em genomas complexos (Alioto *et al.,* 2019). Esta ferramenta utiliza modelos estatísticos de Markov ocultos (HMMs) para predizer a estrutura dos genes, incluindo os locais de início e término, os éxons e os íntrons. O GeneID também leva em consideração características como a frequência de codões e padrões de splicing para melhorar a precisão das predições (Blanco *et al.,* 2007).

Uma das vantagens do GeneID é sua capacidade de adaptar-se a diferentes organismos e condições, tornando-o uma escolha versátil para a anotação de genomas de uma ampla gama de espécies eucarióticas (Legeai *et al.,* 2010). O GeneID já foi aplicado em muitos projetos de anotação de genomas eucarióticos em diferentes organismos, incluindo humanos, animais, plantas e micro-organismos (Blanco *et al.,* 2007). Sua precisão e eficácia na identificação de genes o tornam uma ferramenta valiosa para os pesquisadores que buscam compreender a estrutura e a função dos genomas eucarióticos (Blanco *et al.,* 2009).

O GeneMark é uma ferramenta de predição de genes *ab initio* que é amplamente utilizada na anotação de genomas eucarióticos. Desenvolvida por Alexei Lukashin e Mark Borodovsky, essa ferramenta é notável por sua precisão e eficácia na identificação de genes em

sequências de DNA (Lukashin; Borodovsky, 1998). O GeneMark utiliza modelos estatísticos de Markov ocultos (HMMs) para identificar locais de início e término de genes, bem como regiões de codificação dentro do genoma. Ele é capaz de reconhecer padrões de codões de início, codões de parada e padrões de splicing para melhorar a precisão das predições (Besemer; Borodovsky, 2005).

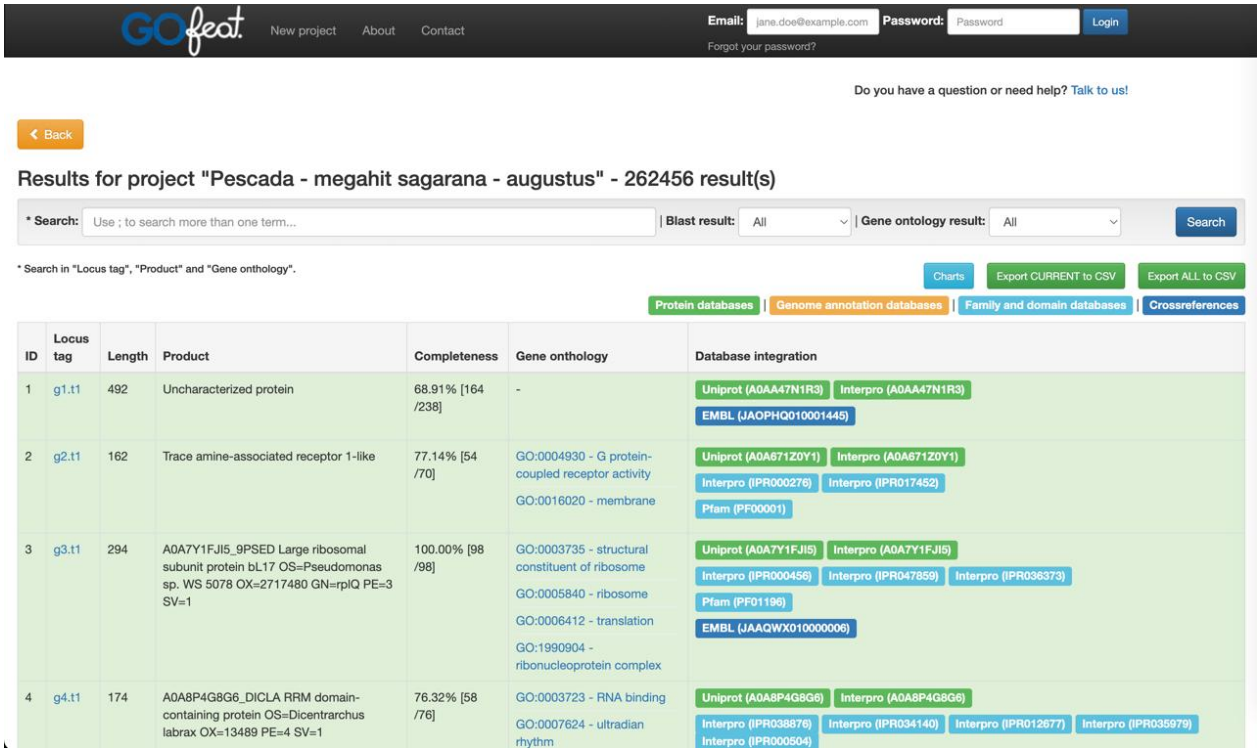
4.5 Anotação do Genoma

Os resultados gerados pelo preditor gênico AUGUSTUS foram usados com entrada no programa GOFEAT (Araujo *et al.,* 2018) por apresentar bons resultados, integração com banco de dados públicos e relatórios detalhados. Que assim gerou a anotação funcional do genoma da pescada amarela.

O processamento do GOFEAT para os dados da pescada amarela foi realizado em um computador com processador Intel(R) Xeon(R) Silver 4210R de 2.40 GHz, 30 cores e 72GB de memória RAM.

A Figura 10 mostra a tela do GOFEAT recebendo os dados do AUGUSTUS para fazer a anotação funcional da pescada amarela.

Figura 10. Tela do Programa GOFEAT produzindo a anotação funcional.



Fonte: Os autores (2024).

Outra ferramenta de bioinformática capaz de realizar a anotação funcional é o EVidenceModeler (EVM), se trata de uma ferramenta de integração de evidências projetada para compilar e refinar previsões de anotação de genes em genomas. Desenvolvido por Mark Yandell e colegas, o EVM permite combinar e pesar diferentes fontes de evidência para produzir uma anotação de gene final mais precisa e abrangente (Haas *et al.,* 2008). Ele é comumente usado em projetos de anotação de genomas, nos quais várias ferramentas de predição de genes, como AUGUSTUS, GeneMark, Glimmer, entre outras, são usadas para gerar previsões de genes (Gabriel *et al.,* 2021) Essas previsões são então combinadas com evidências experimentais, como sequências de ESTs (Expressed Sequence Tags), alinhamentos de proteínas homólogas e dados de RNA-seq, para refinar e melhorar a anotação de genes (Haas *et al.,* 2008).

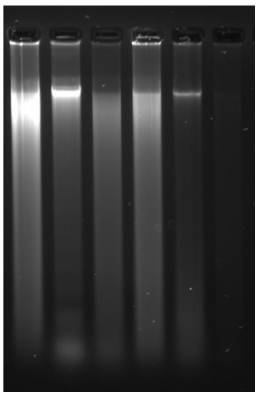
O EVM usa um modelo de aprendizado de máquina para integrar essas diferentes fontes de evidência e produzir uma anotação final que otimiza a precisão e a sensibilidade na identificação de genes (Tang *et al.,* 2014). Ele também permite aos usuários ajustar e personalizar parâmetros para atender às necessidades específicas de seus projetos. Essa abordagem de integração de evidências ajuda a mitigar os erros inerentes a qualquer método individual de predição de genes e a produzir uma anotação final mais confiável e completa (Tang *et al.,* 2014). Portanto, o EVM desempenha um papel importante na anotação de genomas e na compreensão da estrutura e função dos genomas de uma variedade de organismos (Kong *et al.,* 2016).

5. RESULTADOS

5.1 Coleta de Amostras

Dentre as seis amostras de DNA coletadas, a de melhor qualidade foi a de número 5, pois possuía o DNA menos desintegrado e, portanto, a melhor visibilidade e dosagem. Foi então feito o procedimento de eletroforese em gel de agarose em laboratório (Figura 11).

Figura 11. Integridade das amostras de DNA coletadas, feito através do gel de agarose.



	Amostra	Dosagem ng/ul	Razão 260/280
1	2-Pam	297,9	2,04
2	S-Fig	217,9	1,81
3	S-Mus	126,0	1,80
4	C-Mus	235,7	1,97
5	PA VP 3	75,5	1,73
6	PA VP 4	19,2	1,91

Fonte: Os autores (2024).

5.2 Sequenciamento do Genoma

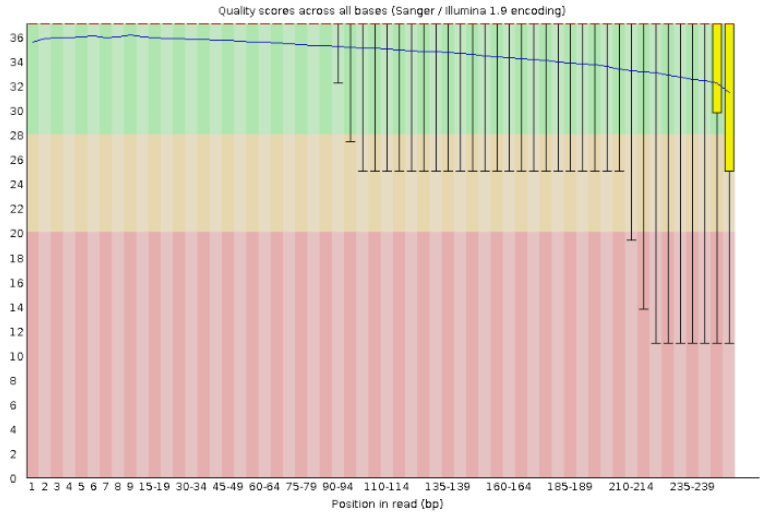
Foi realizado o sequenciamento do genoma completo dos indivíduos capturados com duas bibliotecas pareadas *paired-end short insert* (2 x 250 pb). Cada uma das bibliotecas gerou 2 arquivos, as fitas *Reverse* (Figuras 12 e 14) e *Forward* (Figuras 13 e 15), gerando assim no total 1,302 Gb de sequências de dados brutos. Pode-se observar que ambas as fitas de ambas as bibliotecas obtiveram bons resultados, se mantendo acima do corte *phred* 28 (Figuras 12 a 15). Nos relatórios gerados pelo FastQC, as estatísticas básicas fornecem informações valiosas sobre a qualidade e características dos dados de sequenciamento, incluindo o número total de sequências, o comprimento das sequências, a distribuição do conteúdo de GC e a ausência de sequências consideradas de baixa qualidade.

Figura 12. Relatórios de avaliação do sequenciamento gerados pelo FastQC – Fita L002_R2_001.

Basic Statistics

Measure	Value
Filename	PA-VP-3_S1_L002_R2_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	380666644
Total Bases	90 Gbp
Sequences flagged as poor quality	0
Sequence length	35-251
%GC	42

Per base sequence quality



Total Sequences: O arquivo contém 380,666,644 sequências, o que indica uma quantidade substancial de dados de sequenciamento.

Total Bases: Com um total de 90 Gbp, isso sugere um alto rendimento de dados.

Sequences flagged as poor quality: Nenhuma sequência foi marcada como de baixa qualidade, o que é um bom indicador da qualidade geral do sequenciamento.

Sequence length: As sequências variam de 35 a 251 bases, indicando uma variação no comprimento dos reads.

%GC: O conteúdo GC é de 42%, que está dentro do intervalo esperado para muitos genomas.

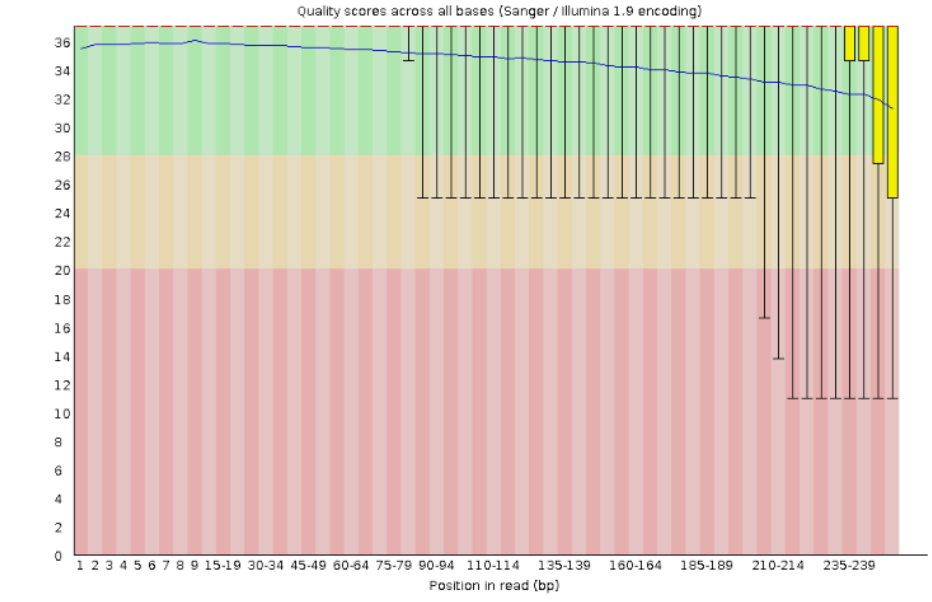
A qualidade da sequência é alta nas primeiras 200 bases, com scores de *Phred* geralmente acima de 28 (faixa verde), indicando alta qualidade. A partir de cerca de 210 bases, a qualidade começa a cair, com algumas leituras entrando na faixa amarela (qualidade moderada) e algumas na faixa vermelha (qualidade baixa). A média de qualidade permanece alta para a maior parte da sequência. Os dados de sequenciamento têm alta qualidade, mas as bases finais mostram uma queda na qualidade, o que é comum em sequenciamentos de leitura longa.

Figura 13. Relatórios de avaliação do sequenciamento gerados pelo FastQC – Fita L001_R2_001.

Basic Statistics

Measure	Value
Filename	PA-VP-3_S1_L001_R2_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	270582401
Total Bases	63.7 Gbp
Sequences flagged as poor quality	0
Sequence length	35-251
%GC	42

Per base sequence quality



A qualidade é alta no início das leituras (pontuações próximas de 36), permanecendo no intervalo verde (alta qualidade) até cerca de 180 bp. A partir de 180 bp, a qualidade começa a

diminuir, caindo para o intervalo amarelo (qualidade moderada) e finalmente para o vermelho (baixa qualidade) após cerca de 220 bp. Existe uma variabilidade maior na qualidade nas posições finais das leituras.

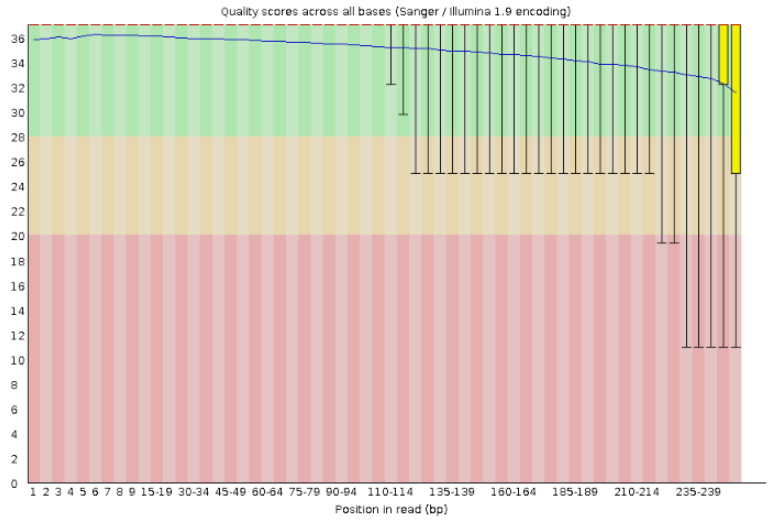
As leituras possuem alta qualidade na maior parte da sequência, com uma diminuição notável na qualidade nas posições finais. No entanto, nenhuma sequência foi marcada como de má qualidade, o que sugere que os dados são confiáveis.

Figura 14. Relatórios de avaliação do sequenciamento gerados pelo FastQC – Fita L002_R1_001.

Basic Statistics

Measure	Value
Filename	PA-VP-3_S1_L002_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	380666644
Total Bases	90.1 Gbp
Sequences flagged as poor quality	0
Sequence length	35-251
%GC	42

Per base sequence quality



As pontuações de qualidade no início das leituras são altas, próximas de 36, e permanecem dentro do intervalo verde (alta qualidade) até cerca de 180 bp. A partir de aproximadamente 180 bp, há uma diminuição gradual na qualidade das bases, movendo-se para o intervalo amarelo (qualidade moderada). Nas posições finais das leituras (após cerca de 220 bp), a qualidade cai significativamente, entrando no intervalo vermelho (baixa qualidade).

A maior parte das sequências tem alta qualidade, especialmente nas posições iniciais e intermediárias das leituras. Há uma diminuição na qualidade das bases nas posições finais das

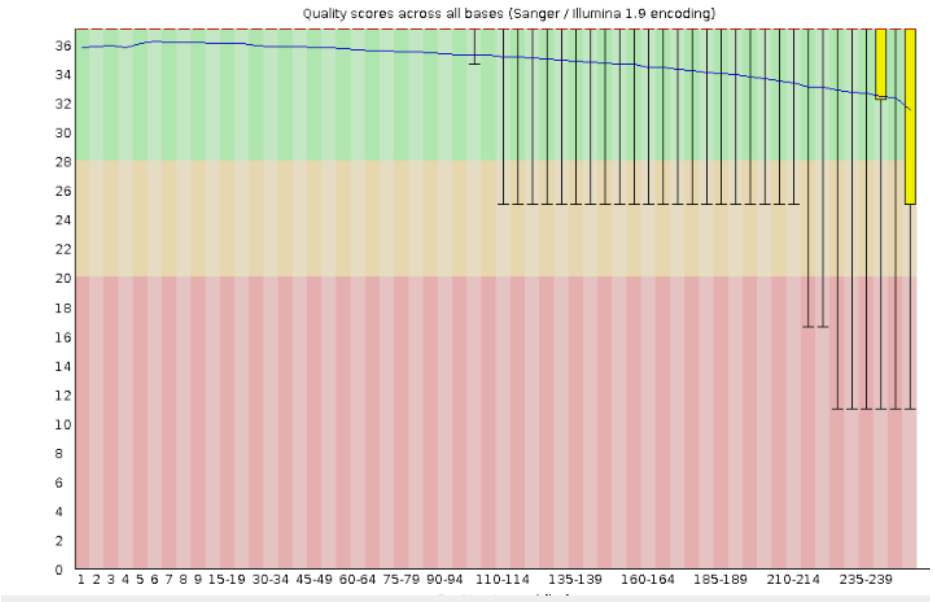
leituras. O que é normal para grandes leituras. Nenhuma sequência foi marcada como de má qualidade, o que indica que os dados são, em geral, confiáveis.

Figura 15. Relatórios de avaliação do sequenciamento gerados pelo FastQC – Fita L001_R1_001.

Basic Statistics

Measure	Value
Filename	PA-VP-3_S1_L001_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	270582401
Total Bases	63.7 Gbp
Sequences flagged as poor quality	0
Sequence length	35-251
%GC	42

Qualidade por sequência base



O gráfico de qualidade por base (Per base sequence quality) mostra a qualidade média das bases em cada posição ao longo das sequências de DNA. Geralmente, a qualidade de uma base é representada pela pontuação de qualidade de Phred, que é um valor numérico que indica a confiabilidade da base. Pontuações de qualidade mais altas indicam bases mais confiáveis e, portanto, uma menor probabilidade de erro de leitura.

Os resultados do FastQC foram positivos, pois todos os gráficos gerados para as fitas *Reverse* e *Forward* apresentaram valores elevados, com a maioria dos resultados situando-se

na faixa verde. Isso sugere uma alta qualidade nos dados de sequenciamento, indicando que as bases estão consistentemente bem distribuídas e com boa qualidade ao longo de ambas as fitas. A predominância de resultados na faixa verde confirma a confiabilidade dos dados, o que é essencial para análises subsequentes de genômica e bioinformática.

5.2 Montagem do Genoma

A ferramenta de Bioinformática QUAST permite avaliar os principais indicadores de qualidade da montagem realizada pelo MEGAHIT. Os resultados são apresentados a seguir.

Figura 16. Relatório geral da montagem da *Cynoscion acoupa* com o MEGAHIT.

Report

	pescada_mega.contigs
# contigs (>= 0 bp)	1461557
# contigs (>= 1000 bp)	212302
# contigs (>= 5000 bp)	7789
# contigs (>= 10000 bp)	257
# contigs (>= 25000 bp)	3
# contigs (>= 50000 bp)	2
Total length (>= 0 bp)	858922370
Total length (>= 1000 bp)	455039030
Total length (>= 5000 bp)	50223340
Total length (>= 10000 bp)	3235389
Total length (>= 25000 bp)	289662
Total length (>= 50000 bp)	262946
# contigs	382736
Largest contig	150291
Total length	575769596
GC (%)	42.13
N50	1911
N90	720
auN	2475.9
L50	91413
L90	286658
# N's per 100 kbp	0.00

All statistics are based on contigs of size >= 500 bp, unless otherwise noted
(e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

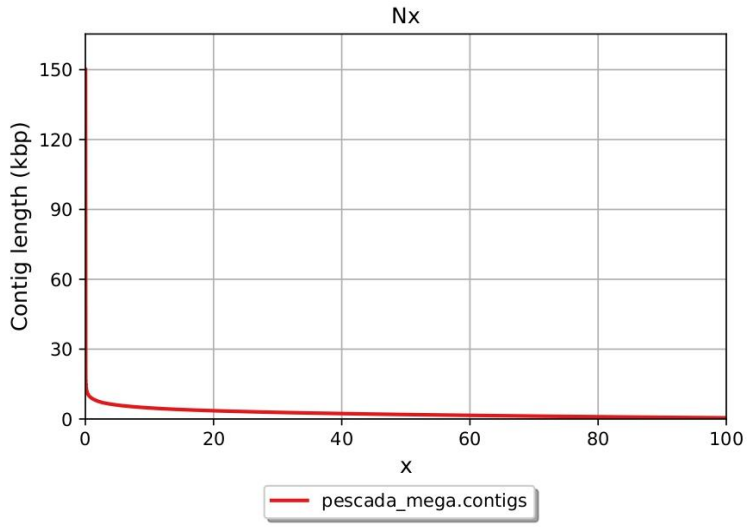
Foram identificados um total de 1.461.557 contigs na amostra. Desses, 212.302 contigs têm pelo menos 1000 base pares (bp), 7.789 têm pelo menos 5000 bp, 257 têm pelo menos 10.000 bp, 3 têm pelo menos 25.000 bp e 2 têm pelo menos 50.000 bp.

O comprimento total dos contigs na amostra é de 858.922.370 bp. Destes, 455.039.030 bp correspondem a contigs com pelo menos 1000 bp, 50.223.340 bp correspondem a contigs com pelo menos 5000 bp, 3.235.389 bp correspondem a contigs com pelo menos 10.000 bp, 289.662 bp correspondem a contigs com pelo menos 25.000 bp e 262.946 bp correspondem a contigs com pelo menos 50.000 bp. O maior contig na amostra tem 150.291 bp de comprimento.

A porcentagem de bases de guanina e citosina (GC) na amostra é de 42,13%. A estatística N50 indica o comprimento em que 50% do comprimento total dos contigs é representado pelos N contigs mais longos. Nesse caso, o N50 é de 1911 bp, enquanto o N90 é de 720 bp. L50 e L90 indicam o número de contigs necessários para alcançar o N50 e o N90, respectivamente. Aqui, são necessários 91.413 contigs para atingir o N50 e 286.658 contigs para atingir o N90. E por fim não foram identificados N's (nucleotídeos indefinidos) a cada 100.000 bp na amostra.

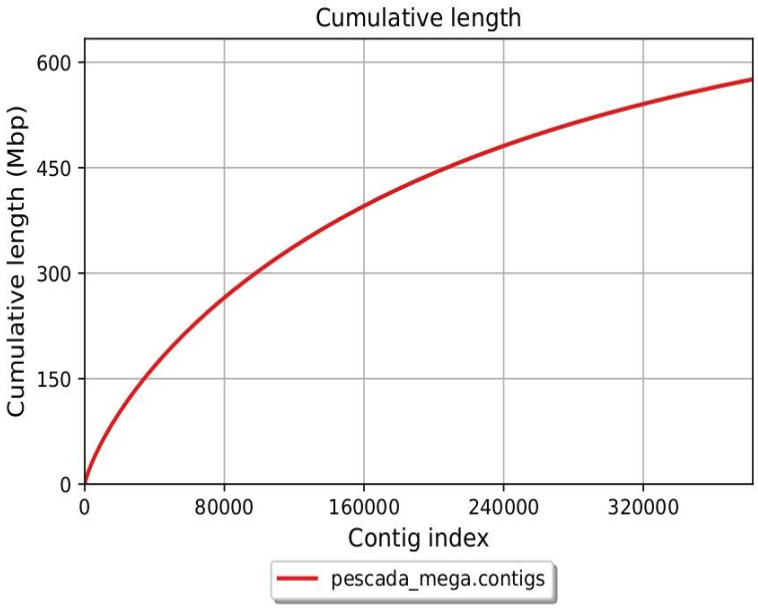
Esses dados fornecem uma visão abrangente da distribuição do tamanho dos contigs, bem como da qualidade geral do sequenciamento genético da amostra.

Figura 17. Comprimento dos contigs após a montagem da *Cynoscion acoupa* com o MEGAHIT.



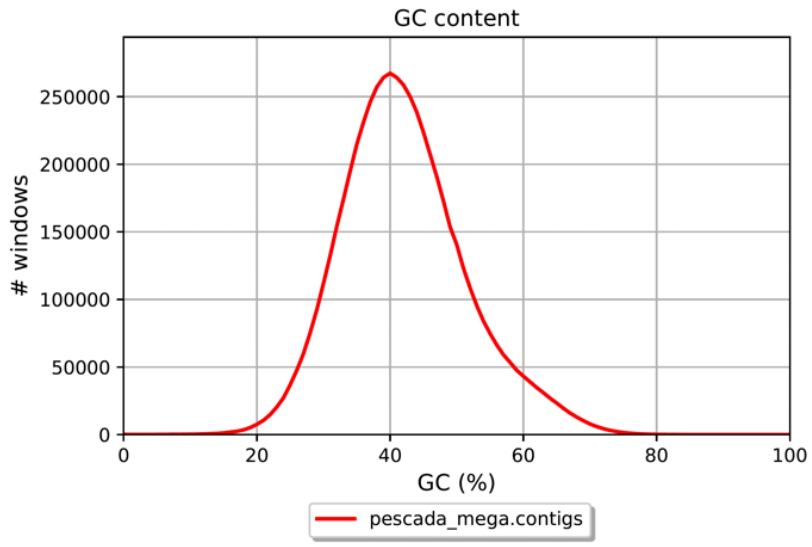
O eixo y representa o comprimento dos contigs (fragmentos de sequências de DNA) em quilobases (kbp) e o eixo x representa o número de contigs. Cada barra ou ponto no gráfico representa um intervalo de comprimento de contig e a altura ou posição da barra ou ponto indica quantos contigs têm esse comprimento. Por exemplo, se houver uma barra ou ponto em x=20 e y=30, isso indicaria que existem 30 contigs no conjunto de dados que têm um comprimento de aproximadamente 20 quilobases. Ele pode fornecer insights sobre a qualidade da montagem, a presença de regiões repetitivas no genoma e outras características estruturais importantes.

Figura 18. Comprimento acumulado das sequências de bases (em megabases) após a montagem da *Cynoscion acoupa* com o MEGAHIT.



A linha curva que passa pelo gráfico sugere uma relação entre o índice de contig e o comprimento acumulado das sequências de bases (em megabases). A curva indica como o comprimento total das sequências de bases aumenta à medida que mais contigs são adicionados. indicando o progresso do sequenciamento genético. A forma da curva pode fornecer insights sobre a distribuição e o tamanho dos contigs na amostra, bem como a cobertura do genoma ou do conjunto de dados sequenciados.

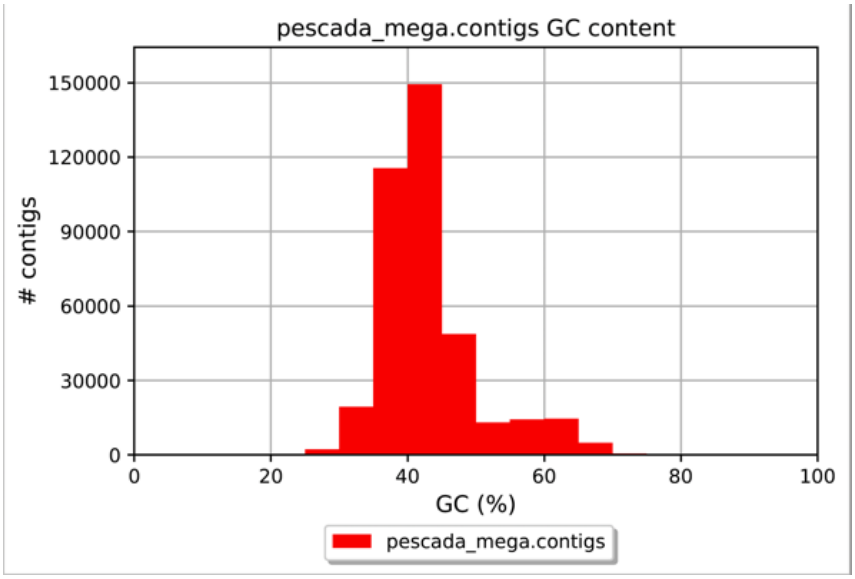
Figura 16. Conteúdo GC x janelas após a montagem.



O conteúdo de GC varia em diferentes regiões ou janelas do genoma ou do conjunto de dados analisado. As regiões com maior conteúdo de GC terão valores mais altos no eixo vertical, que como podemos observar no gráfico é GC 40 com Windows 250000, enquanto as regiões com menor conteúdo de GC terão valores mais baixos. O gráfico também mostra o número de janelas ou regiões analisadas, indicado pela marca "#windows" (número de janelas). Isso sugere que a distribuição do conteúdo de GC foi analisada em várias regiões ou segmentos do genoma ou conjunto de dados.

Esse tipo de análise do conteúdo de GC é importante em genômica e bioinformática, pois o conteúdo de GC pode influenciar vários processos biológicos, como a estabilidade do DNA, a estrutura da cromatina e até mesmo a expressão gênica.

Figura 19. Conteúdo GC x Contigs após a montagem.



O eixo x (horizontal) representa a porcentagem de GC, que é a proporção de bases nitrogenadas guanina e citosina em relação ao total de bases nitrogenadas em uma sequência de DNA. O eixo Y (vertical) representam o número de contigs, que são segmentos de DNA em um genoma que foram montados. O gráfico mostra aumento a partir do GC 30, tendo seu ápice no 40, com 150000 contigs.

Isso sugere que existe uma relação entre a porcentagem de GC e o número de contigs. Picos ou padrões interessantes nesses pontos podem indicar regiões genômicas de interesse, como regiões com características de alta ou baixa GC, ou podem refletir aspectos do processo de montagem do genoma.

5.3 Predição Gênica e Anotação (AUGUSTUS e GOFEAT)

A predição gênica realizada no software AUGUSTUS gerou um total de 262.456 resultados (Figura 18), que foram usados com entrada para o programa **GOFEAT** para gerar a anotação funcional da pescada amarela. A Figura 20 mostra um recorte da saída produzida pelo programa GOFEAT com a anotação funcional do genoma da pescada amarela.

Figura 20. Recorte da Predição Gênica realizada pelo programa AUGUSTUS.

```
>g1.t1
SSPCCATFALQKHVKDSVPGNEDILQSIEQSFYVDNCLQSLSTAESAKLLVDKMRQCLASGGFEIRQWASNLPSVVSHLPSSARSESTELWLTEKSNDPQ
EQALGLRWHCPTDQLGYKSKPPESEIPTMRYIYKVLACQYDPLGIIIPYTTRAKILVQRLWAKKRSWDDTNLPPEILQAWSSWERELPELANILIPRTYA
PMELATDTTEYTLHVFCDSSEQAYGSVAYLTTESDD
>g2.t1
AARAQQLQLSLVMETLEENELCLSQLLNTSCRKPITTQFESMLAYILLSFIALLTSALNLLVIISISHFK
>g3.t1
MRHRKSGRHLSRTSSHRKAMFQNMAVSLFEHELIKTTLPKAKELRRVAEPLITLAKIDSVANRRLAFDRTRSKAMVGKLFNDLGKRYATREGGYLRIL
>g4.t1
MFDQSQASLKEENDSLRWQLDAYRNEVELLRKEQIRSSRADDEHTPTHTHSPEVQIQLLQQSMHNMQQETNQKKTL
>g5.t1
DSCPEGSIDLKCKERVTDSDEDEPDGQHGFQPVARSSLPSSSPSVPHSDSTSAKGNGGGGGGGEEGSERKRKRATDGGEEGSE
>g6.t1
MLGLDGSLVFLEHVFWVVSLNTLFILVFAFCPYHIGHFSVVGLGFEDYIKASHFDGLITTILGYILLAGALIVCHGLASLVKFQRSRRLLGVCYIVVKVS
LLVVMEIGLFPLICGWWLDICSLEMFDATLKDREQSFDSAPGTTMFLHWLVGMVYVFYFASFILLLREVT
>g7.t1
MSCYGYPWPEILRCDQYPADHLMCISSITNSTAHTGGRKVPQASCRDCELEEASSSKDTLETFCRSDFVVKLRLTRLKYSPVSLSQFSLASKLDVLKHGP
LLGGQIRSRIELWLERDATCVRNMTRQHPRGGTFLLTGTVQGERLVVNKAYAWQRRDKELMAAARKWKHYRCRS
>g8.t1
LFEARSRYDELCIPMCVIPATISNNVPGTDFSLGADTAVNAAMEGCDKIKQSASGTKRRVFVVETMGGYCGYLATSTGIAVGADAAYIFEDPINIHDLKT
NVEHLTEKMKKDIQRGLVLRYV
>g9.t1
LAIPLRLPKAITTATKREELEVISAMGADFWTSNLNQQAKNKDLDENRSKPYFYVNPITVEESPDKVLNHSSTSKPDIIPNQNISSSLQNGESPQKVSPE
VKGPAKITSNQEIKYKRPPPRPPSLGSGSGMGLLFSSASSHHTSSSVPPAAERKEEGRGGVGEREERKSMSTSPPPSRPPVPMQSRAAPPLPPAPLRRTS
SRKSTDRDAGERRERKGQNPAKRVGGGERGEEKTASKSGLL
>g10.t1
LTAISCSSLVQEGEGSRIGFVSRSSTSQPQDVETEDAHEPAEDFVACSTAESKVDGEANDSGDDWKPDKSDEELKESDSELQSKTTKKQKRRSGVKAAET
KPEDSHDGFSCKVCEALHQSEVTLVKHVWSHTDDAGGLCGVCGDLSESVETLKDHLLSHHKTDDCHICGESFLDILNLNEHVAAHSGERPHQCDVCHEAF
ALKESLENHQKLHKTDKVHSCYTCHKVFELKEQLKVHCRTHGNKKTHLCGVCGKSLSDYRSLSRHKMTHSGERPHSCLICGRRFKLLGTLRQHEKIHTDR
ERSYLCDVCCKMFLTSKQLQIHMRTHTNEKPYHCGECGKGFTTKGPLTIHMRVHTGETPYRCPDCGWCFKRKINLDNHVTVHSGYKPFVCGICGKACARK
TYLTVHMRTHNGERPYKCTLCDKAFTQSHCLKTHMKSHLVAEAAT
>g11.t1
MGLTDLFQASGDFSGMTSEKVLMNWLKHQGTITVNEDGTEAAALTQVGFMPLSSQIRFTVDHPFLFLIYEHRTDCLVFMGRVANPSQS
>g12.t1
SDGQIVLNVSPTAVRQLHMDNEAVSFEGRFGGVPHTLYVPIAAILGVYARENGQGMVFDLEEPLEADDEVEPDDDGPPPGSEPPRPTGRPSLKVVK
>g13.t1
MAPSSILARNNDAWLTTKIKTQMLTDSAIPGSRIKVVTENGIVYMLGLVTKQEAARATSLVQSVSGVQKIVKLFEYID
>g14.t1
MTNLKQELASIEERVAYQAHERARDIQ
>g15.t1
PRDCSEVELSYLKTKQVSGGTNQTGRGGGGGEKIYTIDSDREMSLHSDPIHYRESRGLAADDLDYPEIYSDHSDNYRKCEQPIIHLNSSPLHHGDSDLLP
DPTYSKHYSLKEKNLSLSPHETLDRYKQTHCRSCLSKVTASYPPGGPYTPTVSASASATRSPYNRCEACLHTANLYDITEDQLLTDPLVSPTLHHQQQQQ
PDEMFGLYWPQTDGPHVQKRNRLRLSRQHSFDNIMLEKPKDVDLGRPARSVSLKEKDRFLDSASDSHYANLFGMRPYSGRLFGTGSKSMLFNHNLEESKR
SKSLYPAHGSENPFLSHSLRDDTSSRLVHGRSSSDIYKQLAVASPAAVLGAPPIKTRNDANNLRSSVKSTTSYCSRDGRIANDMYNKEHVMPYSANKTSA
YPAPRGVLNSAQFSNRRVYKKIPSLESDV
>g16.t1
MGTEDVAARHGGIPGGQRRHSKLVQRSAACYSDKYGHRDTWDCKRREREKGDEWLNSTERSLQQELAVT
```

Figura 21. Recorte da Anotação Funcional da pescada amarela realizada pelo programa GOFEAT.

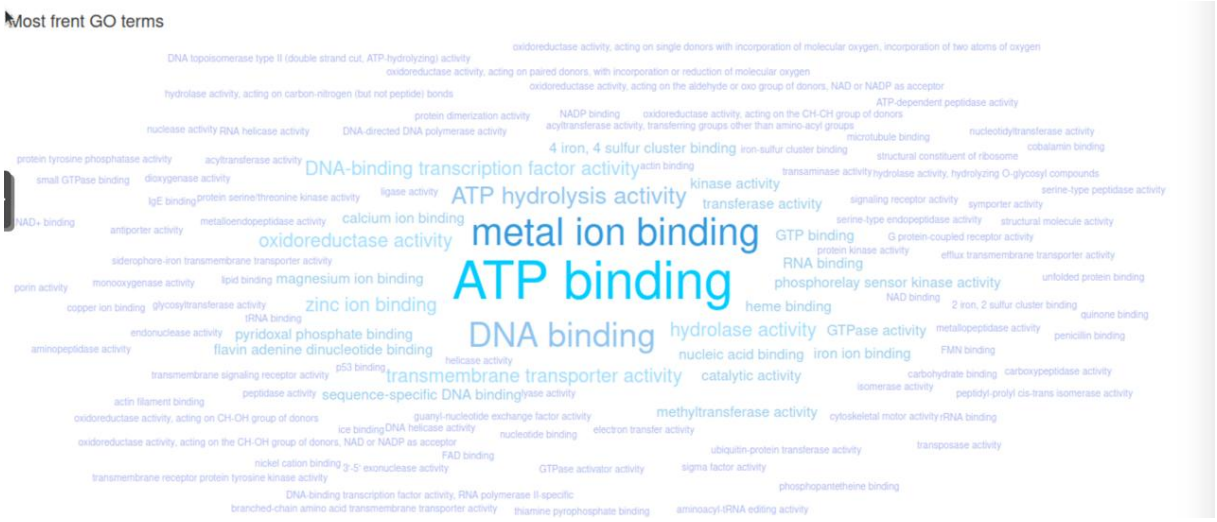
#	Locus tag	Length	Product
1	g1.t1	492	Uncharacterized protein
2	g2.t1	162	Trace amine-associated receptor 1-like
3	g3.t1	294	A0A7Y1FJI5_9PSED Large ribosomal subunit protein bL17 OS=Pseudomonas sp. WS 5078 OX=2717480 GN=rplQ PE=3 SV=1
4	g4.t1	174	A0A8P4G8G6_DICLA RRM domain-containing protein OS=Dicentrarchus labrax OX=13489 PE=4 SV=1
5	g5.t1	240	Protein capicua homolog
6	g6.t1	480	A0A1A8U838_NOTFU RING-type E3 ubiquitin transferase (Fragment) OS=Nothobranchius furzeri OX=105023 GN=Nfu_g_1_020383 PE=4 SV=1
7	g7.t1	486	A0A3B5AZK6_9TELE FZ domain-containing protein OS=Stegastes partitus OX=144197 PE=3 SV=1
8	g8.t1	339	ATP-dependent 6-phosphofructokinase, liver type
9	g9.t1	690	Ras and Rab interactor 3 Ras interaction/interference protein 3
10	g10.t1	1236	A0A6G0HJY3_LARCR C2H2-type domain-containing protein OS=Larimichthys crocea OX=215358 GN=D5F01_LYC23103 PE=4 SV=1
11	g11.t1	237	HEP2
12	g12.t1	288	ClpXP protease specificity-enhancing factor
13	g13.t1	234	Phospholipid-binding protein
14	g14.t1	81	Transmembrane and coiled-coil domain protein 3-like
15	g15.t1	1275	A0A6G0HV01_LARCR Glutamate receptor OS=Larimichthys crocea OX=215358 GN=D5F01_LYC18405 PE=3 SV=1
16	g16.t1	69	Arsb protein
17	g17.t1	111	(spotted green pufferfish) hypothetical protein
18	g18.t1	1962	Tyrosine-protein phosphatase non-receptor type 23
19	g19.t1	627	Inactive dipeptidyl peptidase 10 Dipeptidyl peptidase IV-related protein 3
20	g20.t1	666	Putative methyltransferase
21	g21.t1	354	TetR family transcriptional regulator
22	g22.t1	1005	Calmodulin-binding transcription activator 1
23	g23.t1	300	A0A4W6DVW1_LATCA Protein N-terminal glutamine amidohydrolase OS=Lates calcarifer OX=8187 PE=3 SV=1
24	g24.t1	492	C2 calcium-dependent domain-containing protein 4C
25	g25.t1	228	GON-4-like protein
26	g26.t1	621	Serine/threonine-protein kinase WNK2
27	g27.t1	108	A0A3B4XJR4_SERLL High mobility group AT-hook 2 OS=Seriola lalandi dorsalis OX=1841481 PE=3 SV=1
28	g28.t1	120	Forkhead box protein P4
29	g29.t1	132	Trafficking protein particle complex 9
30	g30.t1	297	ATP-dependent RNA helicase TDRD9
31	g31.t1	159	StAR-related lipid transfer protein 9 START domain-containing protein 9
32	g32.t1	348	A0A4U5VF93_COLLU protein-tyrosine-phosphatase OS=Collichthys lucidus OX=240159 GN=D9C73_020867 PE=3 SV=1
33	g33.t1	240	A0A6G0HM29_LARCR Receptor activity-modifying protein 3 OS=Larimichthys crocea OX=215358 GN=D5F01_LYC22162 PE=3 SV=1
34	g34.t1	129	A0A3B4AZN8_9GOBI Pyroglutamyl-peptidase I OS=Periophthalmus magnuspinnatus OX=409849 PE=3 SV=1
35	g35.t1	264	PAS domain S-box-containing protein/diguanylate cyclase (GGDEF)-like protein
36	g36.t1	828	Sarcalumenin
37	g37.t1	228	Protein-lysine methyltransferase METTL21C
38	g38.t1	279	A0A1A8B6G9_NOTFU LIM/homeobox protein Lhx5 (Fragment) OS=Nothobranchius furzeri OX=105023 GN=LHX5 PE=4 SV=1
39	g39.t1	585	A0A4U5VRX2_COLLU ZW10 interactor OS=Collichthys lucidus OX=240159 GN=D9C73_026603 PE=4 SV=1
40	g40.t1	348	Laminin subunit beta-2-like
41	g41.t1	1002	Mis18-binding protein 1 Kinetochore-associated protein KNL-2-like protein
42	g42.t1	234	A0A0S7FW53_9TELE Myosin-6 (Fragment) OS=Poeciliopsis prolifica OX=188132 GN=MYH3 PE=4 SV=1
43	g43.t1	216	Cholinergic receptor, nicotinic, alpha polypeptide 1
44	g44.t1	408	XPO7
45	g45.t1	87	Uncharacterized protein

Os resultados variam em localização, tamanho e função. Entre os produtos gerados, observou-se uma diversidade de proteínas, incluindo proteínas não caracterizadas, receptores de traço associados a aminas, proteínas ribossomais e proteínas envolvidas em processos metabólicos e de sinalização.

Uma média das proteínas revelou uma variedade de tamanhos e funções, destacando a complexidade e diversidade do genoma estudado. Esses resultados fornecem uma visão abrangente da diversidade genética e do potencial funcional do organismo em questão, abrindo caminho para estudos futuros sobre a regulação genética e as vias biológicas envolvidas.

Ao final do processo de anotação funcional, a ferramenta GOFEAT permite gerar uma nuvem de palavras com as ontologias relacionadas aos produtos encontrados. A nuvem de palavras é exibida na Figura 22.

Figura 22. Ontologias relacionadas aos produtos encontrados após a anotação funcional da pescada amarela com o programa GOFEAT.



A totalidade dos produtos encontrados na anotação funcional da pescada amarela pode ser encontrada no link a seguir: < https://docs.google.com/spreadsheets/d/1qDzZA-7Fx-j9zz9sWMHUMZ3WlE1R5FOk/edit?usp=sharing&ouid=107147507373473858882&rtpof=true&sd=true >

6. DISCUSSÃO

A montagem *de novo* do genoma de um eucarioto é consideravelmente mais complexa e desafiadora em comparação com a montagem de genomas procarióticos. Trata-se de uma tarefa árdua devido à complexidade inerente aos genomas desses organismos, que são geralmente maiores e mais complexos, contendo sequências repetitivas e regiões não codificantes que podem dificultar a montagem correta.

A montagem e anotação genômica de *Cynoscion acoupa* podem ser comparadas com estudos similares realizados em outras espécies de peixes. Por exemplo, a montagem do genoma do *Danio rerio* (peixe-zebra), que é um modelo importante na pesquisa biológica, utilizou ferramentas como Gnomon, GeneMark e EVidenceModeler (EVM), o estudo de Howe *et al.* (2013) sobre o genoma do *Danio rerio* revelou a complexidade do genoma de vertebrados e forneceu insights significativos sobre a evolução dos genes e suas funções em organismos aquáticos. Estudos com o *Danio rerio* feito pelo pesquisador Howe *et al* (2013) têm mostrado a eficiência do uso de ferramentas como o AUGUSTUS na predição gênica devido ao seu genoma bem caracterizado, o que facilita a comparação e validação de dados genômicos entre espécies. Também pode-se observar o estudo do genoma do *Labeo rohita* (carpa-rohu) que revelou informações sobre fluxo gênico e determinação do sexo, utilizando abordagens semelhantes de sequenciamento e anotação no trabalho de Jena *et al.* (2017).

A montagem *de novo* do genoma da pescada amarela foi um trabalho desafiador de diversos aspectos. O primeiro ponto de limitação foram os recursos computacionais. Os algoritmos que realizam montagem *de novo* baseadas em k-mers, ou seja, com a abordagem de grafos *de bruijn*, requerem um alto poder de processamento e quantidade de memória. Quando este processo é aplicado a um eucarioto, em especial um peixe, essa complexidade aumenta exponencialmente. Outro aspecto foi a busca por organismos filogeneticamente próximos para servirem de base durante o processo de predição gênica e anotação funcional. Como não havia genomas próximos, a complexidade do processo aumentou.

As etapas predição gênica realizadas com o software Augustus e a anotação funcional com o GOFEAT revelaram informações inéditas sobre a estrutura e função dos genes da *C. acoupa*. A identificação de sequências codificadoras de proteínas, elementos regulatórios e RNAs não codificantes é crucial para compreender os processos biológicos desta espécie. Esse trabalho é consistente com estudos anteriores, como os de Giani *et al.* (2020) e Charllis *et al.*

(2020), que destacam a importância da montagem genômica na compreensão da estrutura e função dos genomas.

A predição gênica em *Cynoscion acoupa* revelou uma variedade de proteínas envolvidas em processos metabólicos e de sinalização, semelhantes às encontradas em outros peixes de importância econômica, como o salmão (*Salmo salar*) e o bacalhau (*Gadus morhua*). Estudos em salmão têm utilizado ferramentas de predição gênica para identificar genes relacionados à resistência a doenças e crescimento rápido, aspectos cruciais para a aquicultura (Lien *et al.,* 2016)

Os resultados obtidos neste estudo estão alinhados com pesquisas anteriores sobre genômica de peixes, como as de Crepaldi (2024) que trabalhou com a arquitetura genômica de peixes *Megaleporinus*, destacando assim a importância da montagem genômica na compreensão da estrutura genética e na identificação de variações genéticas. No entanto, este estudo é pioneiro na aplicação dessas técnicas à *C. acoupa*, preenchendo uma lacuna na literatura genômica de peixes marinhos-estuarinos.

As ferramentas de bioinformática são amplamente aplicadas em uma variedade de eucariotos, proporcionando uma anotação genômica precisa e detalhada, a anotação genômica do *Cynoscion nebulosus* (pescada-manchada), uma espécie filogeneticamente próxima, revelou importantes informações sobre a estrutura genômica e a evolução de características adaptativas em ambientes estuarinos (Roberts *et al.,* 2009). Os resultados obtidos com a *Cynoscion acoupa* e outros organismos filogeneticamente próximos permite uma análise mais ampla das funcionalidades genéticas, identificando regiões conservadas e elementos regulatórios que podem ser cruciais para a adaptação e sobrevivência dos organismos.

A anotação funcional do genoma de *C. acoupa* tem importantes implicações práticas, especialmente para a indústria pesqueira e farmacêutica (Noda *et al.,* 2010). A bexiga natatória da pescada-amarela, rica em colágeno, é utilizada na produção de *isinglass,* um material valioso para vários setores. O conhecimento detalhado do genoma pode levar ao desenvolvimento de métodos mais eficientes de remoção e utilização deste recurso, aumentando o valor econômico da espécie (Ledur *et al.,* 2004). A anotação funcional feita por programas como o GOFEAT tem sido aplicada para identificar funções genéticas e associações com processos biológicos específicos. Em organismos como a tilápia (*Oreochromis niloticus*), ela tem revelado genes associados a mecanismos de resistência ao estresse ambiental e reprodução, oferecendo *insights* valiosos para a aquicultura sustentável.

A gestão sustentável das populações de *C. acoupa* é crucial para evitar a sobrepesca e garantir a sustentabilidade da indústria pesqueira. O conhecimento genômico pode informar

práticas de manejo e conservação mais eficazes, contribuindo para a preservação dos ecossistemas aquáticos. Regulamentações baseadas em dados genômicos podem promover a pesca sustentável e a conservação da natureza natural de *C. acoupa* (Castro *et al.*, 2018). Além disso, a integração de dados transcriptômicos e proteômicos pode oferecer uma visão mais completa dos processos biológicos em *C. acoupa*.

Como pode se observar nos estudos de Li *et al.* (2016). A colaboração interdisciplinar entre bioinformática, biotecnologia e ecologia será fundamental para explorar plenamente o potencial genômico desta espécie. A montagem genômica em *Cynoscion acoupa* e outros organismos aquáticos proporciona uma base sólida para pesquisas futuras, incluindo a aplicação de técnicas avançadas como a edição de genes CRISPR-Cas9. Essa técnica tem sido explorada em organismos como o peixe-zebra para entender a função de genes específicos e desenvolver novas abordagens terapêuticas (Hoshijima *et al.,* 2016).

7. CONSIDERAÇÕES FINAIS

A montagem *de novo*, no contexto da genômica, é uma técnica fundamental para criar sequências genômicas completas de organismos, especialmente quando não há um genoma de referência disponível ou quando se deseja obter uma visão completa e não tendenciosa do genoma. Existem várias razões para usar a montagem *de novo*, como, a descoberta de novas espécies; a variação genica; identificação de novos genes e elementos regulatórios. Vale ressaltar que não até o momento não existia nenhum genoma sequenciado e disponível, dos organismos pertencentes ao gênero *Cynoscion*, que é o gênero onde se encontra o pescado *C. acoupa.*

O projeto de montagem, predição gênica e anotação do genoma da espécie *Cynoscion acoupa* apresentou resultados promissores e inéditos. Através da coleta de amostras nas regiões de Salinópolis-PA, Bragança-PA e na costa do Amapá, e utilizando a tecnologia de sequenciamento NovaSeq SP 6000 Illumina, foi possível gerar aproximadamente 1302 Gb de dados brutos. A qualidade do sequenciamento foi verificada utilizando o FASTQC, e a montagem *de novo* do genoma foi realizada com o montador MEGAHIT. Foi tentada a execução dessa montagem no servidor local que estava disponível na UFRA campus Paragominas, mas ele não foi capaz de finalizar a montagem. Posteriormente, tentou-se executar a montagem em um servidor mais robusto de uma instituição parceira na República Dominicana. O processo também não conseguiu finalizar. Por fim, conseguiu-se realizar a montagem em um terceiro servidor da Universidade Federal de Minas Gerais.

A predição gênica, executada com a ferramenta AUGUSTUS, e a anotação funcional, realizada com o software GOFEAT, resultaram em uma compreensão detalhada e abrangente do genoma da *Cynoscion acoupa*. Os resultados obtidos forneceram informações valiosas sobre a complexidade e diversidade do genoma desta espécie, destacando a presença de uma variedade de proteínas, incluindo proteínas não caracterizadas, receptores associados a aminas, proteínas ribossomais e proteínas envolvidas em processos metabólicos e de sinalização.

Como trabalho futuro, é preciso realizar o depósito do genoma no NCBI, o que contribuirá significativamente para futuras pesquisas, permitindo que outros pesquisadores tenham acesso a esses dados para estudos complementares. Este projeto não apenas enriquece o conhecimento científico sobre a *Cynoscion acoupa*, mas também abre novas possibilidades de aplicação biotecnológica, especialmente na indústria farmacêutica e alimentícia. Em suma, este estudo pioneiro na montagem, predição gênica e anotação do genoma da *Cynoscion acoupa*

representa um avanço significativo na genômica de peixes de importância econômica, oferecendo uma base sólida para futuras investigações e aplicações práticas.

As tarefas de sequenciamento, montagem, predição gênica e anotação funcional do genoma da *Cynoscion acoupa,* conhecida como pescada amarela, representam avanços importantes no campo da biotecnologia e bioinformática na Amazônia, principalmente por se tratar de um estudo pioneiro, a primeira caracterização genômica desta espécie, oferece uma base robusta para futuras pesquisas genômicas.

REFERÊNCIAS

Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi E Glenn Tesler,
Quast: Ferramenta De Avaliação De Qualidade Para Montagens De Genoma,
Bioinformática (2013) 29 (8): 1072-1075. Doi: 10.1093/Bioinformatics/Btt086
Publicado Pela Primeira Vez Online: 19 De fevereiro De 2013.

Alioto, Tyler, Et Al. Using Geneid To Identify Genes. *Current Protocols In Bioinformatics*, 2018, 64.1: E56.

Alkan, Can; Sajjadian, Saba; Eichler, Evan E. Limitations Of Next-Generation Genome Sequence Assembly. Nature Methods, 2011, 8.1: 61-65.

Andrews, S. *Et Al.,* (2004) *Fastqc*, *Babraham Bioinformátics: Fastqc*. Available At: Https://Www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc/ (Accessed: 26 June 2023).

Araujo, F., Barh, D., Silva, A. *Et Al.* Go Feat: A Rapid Web-Based Functional Annotation Tool For Genomic And Transcriptomic Data. *Sci Rep* **8**, 1794 (2018). Https://Doi.Org/10.1038/S41598-018-20211-9

Arick, Mark A., Et Al. A High-Quality Chromosome-Level Genome Assembly Of Rohu Carp, Labeo Rohita, And Its Utilization In Snp-Based Exploration Of Gene Flow And Sex Determination. *G3: Genes, Genomes, Genetics*, 2023.

Ayulo, Andrés Mané Romero; Machado, Rubén Abreu; Scussel, Vildes Maria. Escherichia Coli E Staphylococcus Aureus Enterotoxigênicas Em Peixes E Frutos Do Mar Da Região Sul Do Brasil. Jornal Internacional De Microbiologia Alimentar , 1994, 24.1-2: 171-178.

Baudet, Christian. "Uma abordagem para trimagem, verificaçao de contaminaçao e clusterizaçao de seqüências EST." *Master's thesis, Universidade Estadual de Campinas (Unicamp)* 14 (2006): 33.

Bayat, Arash, Et Al. "Fast Short Read De-Novo Assembly Using Overlap-Layout-Consensus Approach." Ieee/Acm Transactions On Computational Biology And Bioinformatics 17.1 (2018): 334-338.

Behjati, S. And Tarpey, P.S. (2013) 'What Is Next Generation Sequencing?', *Archives Of Disease In Childhood - Education & Practice Edition*, 98(6), Pp. 236–238. Doi:10.1136/Archdischild-2013-304340.

Besemer, John; Borodovsky, Mark. Genemark: Web Software For Gene Finding In Prokaryotes, Eukaryotes And Viruses. *Nucleic Acids Research*, 2005, 33.Suppl_2: W451-W454.

Besemer, John; Borodovsky, Mark. Genemark: Web Software For Gene Finding In Prokaryotes, Eukaryotes And Viruses. *Nucleic Acids Research*, 2005.
Birol, Inanç, Et Al. De Novo Transcriptome Assembly With Abyss. *Bioinformatics*, 2009, 25.21: 2872-2877.

Black, D. L. (2000). Protein diversity from alternative splicing: A challenge for bioinformatics and post-genome biology**.** *Cell*, 103(3), 367-370. doi:10.1016/S0092-8674(00)00127-7

Blanco, Enrique; Abril, Josep F. Computational Gene Annotation In New Genome Assemblies Using Geneid. *Bioinformatics For Dna Sequence Analysis*, 2009.

Blanco, Henrique; Parra, Genís; Guigó, Rodérico. Usando Geneid Para Identificar Genes. *Protocolos Atuais Em Bioinformática* , 2007.

Boulhosa, Amanda Mariana. Alternativa Para O Uso Sustentável Da Pescada Amarela Cynoscion Ocoupa (Lacépede, 1802) Capturada Com Rede De Emalhe De Fundo Na Costa Norte Do Brasil. 2012.

Brůna, Tomáš, Et Al. Galba: Genome Annotation With Miniprot And Augustus. *Bmc Bioinformatics*, 2023

Brůna, Tomáš; Lomsadze, Alexandre; Borodovsky, Mark. Genemark-Ep+: Eukaryotic Gene Prediction With Self-Training In The Space Of Genes And Proteins. *Nar Genomics And Bioinformatics*, 2020.

BULGARELLI, Caio et al. Método e sistema para predição de funções de proteínas. 2018.
Burge, C., & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268(1), 78-94.

Carpenter, Eric J., Et Al. Soap-Denovo-Trans Assembly. 2019.

Castro, Isadora Fontenelle Carneiro De, Et Al. Pesquisa De Listeria Monocytogenes E Vibrio Parahaemolyticus Em Amostras De Pescada Amarela (*Cynoscion Acoupa*) Comercializadas Na Cidade De São Luís, Ma. Hig. Aliment, 2018, 99-102.

Chikhi, Rayan; Medvedev, Paul. Informed And Automated K-Mer Size Selection For Genome Assembly. *Bioinformatics*, 2014.

De Almeida, Zafira Da Silva, Et Al. Avaliação Do Potencial De Produção Pesqueira Do Sistema Da Pescada-Amarela (Cynoscion Acoupa) Capturada Pela Frota Comercial Do Araçagi, Raposa, Maranhão. 2011.

De Almeida, Zafira Da Silva, Et Al. Contribuição Para Gestão Do Sistema De Produção Pesqueira Pescada-Amarela, Cynoscion Acoupa (Pisces: Sciaenidae)(Lacepéde, 1802) Na Costa Do Maranhão, Brasil. 2014.

De Keersmaecker, S.C. *Et Al.* (2006) 'Integration Of Omics Data: How Well Does It Work For Bacteria?', *Molecular Microbiology*, 62(5), Pp. 1239–1250. Doi:10.1111/J.1365-2958.2006.05453.X.

De Matos, Igor Penedo; Lucena, Flávia. Descrição Da Pesca Da Pescada-Amarela, Cynoscion Acoupa, Da Costa Do Pará. Arquivos De Ciências Do Mar, 2017, 39.1-2: 66-73.

De Meira Gusmão, Alexandre Oliveira; Da Silva, Antonio Rodrigues; Medeiros, Mauro Osvaldo. A Biotecnologia E Os Avanços Da Sociedade. *Biodiversidade*, 2017, 16.1.

De Sena Brandine, Guilherme; Smith, Andrew D. Falco: High-Speed Fastqc Emulation For Quality Control Of Sequencing Data. *F1000research*, 2019, 8.
Dunn, Nathan A., Et Al. Apollo: Democratizing Genome Annotation. *Plos Computational Biology*, 2019.

Edwards, Yvonne Jk; Cottage, Amanda. Prediction Of Protein Structure And Function By Using Bioinformatics. *Genomics Protocols*, 2001.

Edwards, Yvonne Jk; Cottage, Amanda. Prediction Of Protein Structure And Function By Using Bioinformatics. *Genomics Protocols*, 2001.

El-Metwally, S. *Et Al.* (2013) 'Next-Generation Sequence Assembly: Four Stages Of Data Processing And Computational Challenges', *Plos Computational Biology*, 9(12). Doi:10.1371/Journal.Pcbi.1003345.

Ferreira, Elka Machado, Et Al. Alterações Sensoriais, Microbiológicas E Químicas Da Pescada Amarela (Cynoscion Acoupa) E Do Peixe-Serra (Scomberomorus Brasiliensis) Desembarcados Em Portos No Maranhão. Brazilian Journal Of Development, 2020, 6.5: 26662-26676.

Ferreira, Maurício Alexander De Moura. "Bioinformática Como Ferramenta No Melhoramento Genético De Plantas." 2018.

Firtina, Can, Et Al. Apollo: A Sequencing-Technology-Independent, Scalable And Accurate Assembly Polishing Algorithm. *Bioinformatics*, 2020.

Firtina, Can, Et Al. Apollo: A Sequencing-Technology-Independent, Scalable And Accurate Assembly Polishing Algorithm. *Bioinformatics*, 2020.

Freire, Jeandria Negreiro, Et Al. Aspectos Da Pesca E Análise Da Abundância Relativa Da Cynoscion Acoupa, Lacépède, 1801 E Suas Relações Com A Temperatura Da Superfície Do Mar Na Plataforma Continental Norte Do Brasil. 2019.

Gabler, Felix, Et Al. Protein Sequence Analysis Using The Mpi Bioinformatics Toolkit. *Current Protocols In Bioinformatics*, 2020.

Gabler, Felix, Et Al. Protein Sequence Analysis Using The Mpi Bioinformatics Toolkit. *Current Protocols In Bioinformatics*, 2020.

Gabriel, Lars, Et Al. Tsebra: Transcript Selector For Braker. *Bmc Bioinformatics*, 2021, 22.1: 1-12.

Gandra, A. Consumo de pescado cresce 65% no brasil desde 2004. Agencia Brasil., 2023

Giani, Alice Maria, Et Al. Long Walk To Genomics: History And Current Approaches To

Genome Sequencing And Assembly. Computational And Structural Biotechnology Journal, 2020, 18: 9-19.

Giuffra, Elisabetta; Tuggle, Christopher K.; Faang Consortium. Functional Annotation Of Animal Genomes (Faang): Current Achievements And Roadmap. Annual Review Of Animal Biosciences, 2019, 7: 65-88.
Guimarães, R., 2022. Variabilidade Genética Da Pescada Amarela (*Cynoscion Acoupa* Sciaenidae, Lacepède, 1801) E Percepção Ambiental Dos Pescadores Para Sua Conservação No Litoral Maranhense, Brasil. Mestrado. Universidade Estadual Do Maranhão.

Guimarães, Ricardo Fonseca. Variabilidade Genética Da Pescada Amarela (Cynoscion Acoupa-Sciaenidae, Lacepède, 1801) E Percepção Ambiental Dos Pescadores Para Sua Conservação No Litoral Maranhense, Brasil. 2018. Phd Thesis. Uema.

Gupta, Amit Kumar; Kumar, Manoj. Benchmarking And Assessment Of Eight De Novo Genome Assemblers On Viral Next-Generation Sequencing Data, Including The Sars-Cov-2. Omics: A Journal Of Integrative Biology, 2022, 26.7: 372-381.

Gupta, Sushim Kumar, Et Al. Arg-Annot, A New Bioinformatic Tool To Discover Antibiotic Resistance Genes In Bacterial Genomes. *Antimicrobial Agents And Chemotherapy*, 2014, 58.1: 212-220.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. Quast: Quality Assessment Tool For Genome Assemblies. Bioinformatics. 2013 Apr 15;29(8):1072-5. Doi: 10.1093/Bioinformatics/Btt086. Epub 2013 Feb 19. Pmid: 23422339; Pmcid: Pmc3624806.

Haas, Brian J., Et Al. Automated Eukaryotic Gene Structure Annotation Using Evidencemodeler And The Program To Assemble Spliced Alignments. *Genome Biology*, 2008.

Heikema, Astrid P., Et Al. Comparison Of Illumina Versus Nanopore 16s Rrna Gene Sequencing Of The Human Nasal Microbiota. *Genes*, 2020, 11.9: 1105.

Heller, D. And Vingron, M. (2020) 'Svim-Asm: Structural Variant Detection From Haploid And Diploid Genome Assemblies', *Bioinformatics*, 36(22–23), Pp. 5519–5521. Doi:10.1093/Bioinformatics/Btaa1034.

Hoff, Katharina J., Et Al. Braker1: Unsupervised Rna-Seq-Based Genome Annotation With Genemark-Et And Augustus. *Bioinformatics*, 2016

Hoff, Katharina J., Et Al. Whole-Genome Annotation With Braker. *Gene Prediction: Methods And Protocols*, 2019, 65-95.

Hoshijima, K., et al. (2016). "Highly efficient CRISPR-Cas9-based methods for generating deletion mutations and F0 embryos that lack gene function in zebrafish." Developmental Cell, 36, 654-667.

Howe, K., et al. (2013). "The zebrafish reference genome sequence and its relationship to the human genome." Nature, 496, 498-503.

Howe, Kerstin. The Zebrafish Genome Sequencing Project: Bioinformatics Resources. In: *Behavioral And Neural Genetics Of Zebrafish*. Academic Press, 2020. P. 551-562.

Hu, Taishan, Et Al. Next-Generation Sequencing Technologies: An Overview. *Human Immunology*, 2021, 82.11: 801-811.
Jena, J. K., et al. (2017). "Genome-wide identification and characterization of microRNAs in rohu (Labeo rohita) and their potential role in development and immunity." BMC Genomics, 18, 541.

Jeon, Sol A., Et Al. Comparison Between Mgi And Illumina Sequencing Platforms For Whole Genome Sequencing. *Genes & Genomics*, 2021, 43.7: 713-724.

Jorge, Paulo Henrique. Sequenciamento Do Transcriptoma E Caracterização De Microssatélites Na Pirapitinga Piaractus Brachypomus Para Análises De Variabilidade Genética. 2016.

Kaundal, Rakesh; Kapoor, Amar S.; Raghava, Gajendra Ps. Machine Learning Techniques In Disease Forecasting: A Case Study On Rice Blast Prediction. *Bmc Bioinformatics*, 2006.

Khan, Abdul Rafay, Et Al. A Comprehensive Study Of De Novo Genome Assemblers: Current Challenges And Future Prospective. Evolutionary Bioinformatics, 2018, 14: 1176934318758650.

Khew, Choy Yuen, Et Al. Transcriptional Sequencing And Gene Expression Analysis Of Various Genes In Fruit Development Of Three Different Black Pepper (Piper Nigrum L.) Varieties. International Journal Of Genomics, 2020.

Klasberg, Steffen, Et Al. Bioinformatics Strategies, Challenges, And Opportunities For Next Generation Sequencing-Based Hla Genotyping. *Transfusion Medicine And Hemotherapy*, 2019, 46.5: 312-325.

Kong, Jinhwa, Et Al. Draft Genome Of Toxocara Canis, A Pathogen Responsible For Visceral Larva Migrans. *The Korean Journal Of Parasitology*, 2016.

Kooij, Pepijn W.; Pellicer, Jaume. Genome Size Versus Genome Assemblies: Are The Genomes Truly Expanded In Polyploid Fungal Symbionts?. Genome Biology And Evolution, 2020, 12.12: 2384-2390.

Kremer, F. Pós-montagem de genomas: Parte II - Scaffolding. Medium, 8 jul. 2020. Disponível em: <https://medium.com/omixdata/p%C3%B3s-montagem-de-genomas-parte-ii-scaffolding-7c78a218dfca>. Acesso em: 10 jul. 2024.

Lee, Ed, Et Al. Apollo: A Community Resource For Genome Annotation Editing. *Bioinformatics*, 2009.

Legeai, Fabrice, Et Al. Aphidbase: A Centralized Bioinformatic Resource For Annotation Of The Pea Aphid Genome. *Insect Molecular Biology*, 2010.

Li D, Liu Cm, Luo R, Sadakane K, Lam Tw. Megahit: An Ultra-Fast Single-Node Solution For Large And Complex Metagenomics Assembly Via Succinct De Bruijn Graph.

Bioinformatics. 2015 May 15;31(10):1674-6. Doi: 10.1093/Bioinformatics/Btv033. Epub 2015 Jan 20. Pmid: 25609793.

Li D, Luo R, Liu Cm, Leung Cm, Ting Hf, Sadakane K, Yamashita H, Lam Tw. Megahit V1.0: A Fast And Scalable Metagenome Assembler Driven By Advanced Methodologies And Community Practices. Methods. 2016 Jun 1;102:3-11. Doi: 10.1016/J.Ymeth.2016.02.020. Epub 2016 Mar 21. Pmid: 27012178

Li, Fangping, Et Al. Gap-Free Genome Assembly And Comparative Analysis Reveal The Evolution And Anthocyanin Accumulation Mechanism Of Rhodomyrtus Tomentosa. Horticulture Research, 2023, 10.3: Uhad005.

Li, Fei, Et Al. Genomas De Insetos: Progressos E Desafios. Biologia Molecular De Insetos , 2019, 28,6: 739-758.

Li, Xihao, Et Al. Dynamic Incorporation Of Multiple In Silico Functional Annotations Empowers Rare Variant Association Analysis Of Large Whole-Genome Sequencing Studies At Scale. Nature Genetics, 2020, 52.9: 969-983.

Lien, S., et al. (2016). "The Atlantic salmon genome provides insights into rediploidization." Nature, 533, 200-205.

Lin, Yu-Cheng; Wu, Chi-Chien; Ni, Yen-Hsuan. New Perspectives On Genetic Prediction For Pediatric Metabolic Associated Fatty Liver Disease. Frontiers In Pediatrics, 2020, 8: 603654.

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O., & Borodovsky, M. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research*, 33(20), 6494-6506.

Lu, Hengyun; Giordano, Francesca; Ning, Zemin. Oxford Nanopore Minion Sequencing And Genome Assembly. Genomics, Proteomics & Bioinformatics, 2016, 14.5: 265-279. Lukashin, Alexander V.; Borodovsky, Mark. Genemark. Hmm: New Solutions For Gene Finding. *Nucleic Acids Research*, 1998.

Madeira, Humberto Maciel França. Promessas E Desafios Dos Avanços Biotecnológicos. *Revista Acadêmica Ciência Animal*, 2008, 6.2: 309-316.

Madritsch, Silvia; Burg, Agnes; Sehr, Eva M. Comparing De Novo Transcriptome Assembly Tools In Di-And Autotetraploid Non-Model Plant Species. Bmc Bioinformatics, 2021, 22: 1-17.

Mcininch, James D.; Hayes, William S.; Borodovsky, Mark. Applications Of Genemark In Multispecies Environments. In: *Ismb*. 1996.

Medeiros, Adriano Silva, Et Al. Caracterização Do Processamento E Do Comércio De "Grude" Da Pescada-Amarela Cynoscion Acoupa (Lacépède, 1801) Do Município De Apicum-Açu, No Estado Do Maranhão. 2019.

Mendes, Izabelle Da Silva, Et Al. Análise Morfológica De Escamas De Peixes Teleósteos Do Alto Rio Guamá Na Mesorregião Nordeste Paraense. 2019.

Methods,Genomics, Proteomics & Bioinformatics, Volume 2, Issue 4, 2004, Pages 216-221, Issn 1672-0229, Https://Doi.Org/10.1016/S1672-0229(04)02028-5.

Miao, Hongmei, Et Al. Sesame Genome Assembly. The Sesame Genome, 2021, 225-237.
Mikheenko, Alla, Et Al. Versatile Genome Assembly Evaluation With Quastlg. Bioinformatics, 2018, 34.13: I142-I150.

Monte, Flávia Thuane Duarte Do. Caracterização Do Colágeno Extraído A Partir De Escamas De Pescada Amarela (*Cynoscion Acoupa*). Ms Thesis. Universidade Federal De Pernambuco, 2017.

Montelione, G.T. And Anderson, S. (1999) *Nature Structural Biology*, 6(1), Pp. 11–12. Doi:10.1038/4878.

Moreira, L. M. (2015). Ciências genômicas: fundamentos e aplicações. *Moreira, LM & Varani, AM Plasticidade e fluxo genômico. Ribeirão Preto: Sociedade Brasileira de Genética*, *1*, 101-116.

Moura, Hanna Tereza Garcia De Sousa, Et Al. Indicadores De Desempenho Para A Pesca Em Larga Escala Do Peixe Fraco Acoupa Na Plataforma Continental Amazônica. Gestão E Ecologia Pesqueira , 2023.

Mourão, K. R. M.; Frédou, F. L.; Espirito-Santo, R. V.; Almeida, M. C.; Silva, B. B.; Frédou, T.; Isaac, V. Sistema De Produção Pesqueira Pescada-Amarela – Cynoscion Acoupa Lacépède (1802): Um Estudo De Caso No Litoral Nordeste Do Pará, Brasil. Boletim Do Instituto Da Pesca, São Paulo, V. 35. N. 3, P. 497-511, 2009.

Mourão, Keila Renata Moreira, Et Al. Sistema De Produção Pesqueira Pescada Amarela-Cynoscion Acoupa Lací¨ Pede (1802): Um Estudo De Caso No Litoral Nordeste Do Pará-Brasil. Boletim Do Instituto De Pesca , 2009, 35.3:497-511.

Nagarajan, Niranjan; Pop, Mihai. Sequence Assembly Demystified. Nature Reviews Genetics, V. 14, N. 3, P. 157-167, 2013.

Nagy, Alinda, Et Al. Identification And Correction Of Abnormal, Incomplete And Mispredicted Proteins In Public Databases. *Bmc Bioinformatics*, 2008.
Ncbi Blast. Disponível Em <Https://Www.Ncbi.Nlm.Nih.Gov/Search/All/?Term=Cynoscion%20acoupa>. Acesso Em: Setem. 2023.

Nikolić, Vladimir, Et Al. Rresolver: Efficient Short-Read Repeat Resolution Within Abyss. *Bmc Bioinformatics*, 2022, 23.1: 1-15.

Oehmen, Chris; Nieplocha, Jarek. Scalablast: A Scalable Implementation Of Blast For High-Performance Data-Intensive Bioinformatics Analysis. *Ieee Transactions On Parallel And Distributed Systems*, 2006.

Oliveira, Cícero Diogo, Et Al. Biologia E Pesca De Acoupa Fraquinho *Cynoscion Acoupa* (Lacepède, 1801): Uma Revisão. Biologia Neotropical E Conservação, 2020, 15: 333.

OLIVEIRA, Luan Freitas de et al. Investigações genômicas de distúrbios do desenvolvimento. 2019.

Oliveira-Da-Silva, Fúvio Rubens; Ilkiu-Borges, Anna Luiza. Briófitas (Bryophyta E Marchantiophyta) Das Cangas Da Serra Dos Carajás, Pará, Brasil. Rodriguésia, 2018, 69: 1405-1416.

Oyawoye, O. M., Et Al. Bioinformatics: A Tool For Biotechnological Advancement 1. In: *Medical Biotechnology, Biopharmaceutics, Forensic Science And Bioinformatics*. Crc Press, 2022. P. 213-223.

Paszkiewicz, K. And Studholme, D.J. (2010) 'De Novo Assembly Of Short Sequence Reads', *Briefings In Bioinformatics*, 11(5), Pp. 457–472. Doi:10.1093/Bib/Bbq020.

Peker, N. *Et Al.* (2019) 'A Comparison Of Three Different Bioinformatics Analyses Of The 16s–23s Rrna Encoding Region For Bacterial Identification', *Frontiers In Microbiology*, 10. Doi:10.3389/Fmicb.2019.00620.

Pereira, Glauce Vasconcelos Da Silva, Et Al. Aproveitamento Sustentável Dos Resíduos De Pescado Para Obtenção De Revestimentos/Filmes, Aplicação Na Conservação Pós-Colheita De Goiabas (Psidium Guajava L.), Estudo Da Estabilidade E Funcionalidade Desses Filmes. 2021.

Pinheiro, Leina Maria Herculano Maia. Elaboração E Caracterização De Gelatina De Bexiga Natatória De Robalo (Centropomus Undecimalis). 2021.

Pop, M. And Salzberg, S.L. (2008) 'Bioinformatics Challenges Of New Sequencing Technology', *Trends In Genetics*, 24(3), Pp. 142–149. Doi:10.1016/J.Tig.2007.12.006.

Ramos, Lucas Pansani. *Draft Genomes Comparison With Succinct De Bruijn Graphs= Comparação De Genomas Incompletos Usando Grafos De De Bruijn Sucintos*. 2022. Phd Thesis. [Sn].

Reed, Jennifer L., Et Al. Towards Multidimensional Genome Annotation. *Nature Reviews Genetics*, 2006, 7.2: 130-141.

Richardson, E.J. And Watson, M. (2012) 'The Automatic Annotation Of Bacterial Genomes', *Briefings In Bioinformatics*, 14(1), Pp. 1–12. Doi:10.1093/Bib/Bbs007.

Roberto Brito Xavier Junior Fatores De Transcrição Em Cianobactérias: Predição Por Genômica Comparativa/ Roberto Brito Xavier Junior. – Belém, 2018. 46 P. : Il. (Algumas Color.) ; 30 Cm.

Roberts, S. B., et al. (2009). "Isolation and characterization of microsatellite loci in Cynoscion nebulosus, a candidate for stock enhancement." Conservation Genetics Resources, 1, 105-107.

Rodrigues De Lemos, R. (2009). *Análise In Silico De Novos Potenciais Polimorfismos Genéticos De Risco Na Doença De Alzheimer Em Bancos De Dados De Microarrays* (Master's Thesis, Universidade Federal De Pernambuco).

Rodrigues, Renato Pinheiro, Et Al. A Pesca Esportiva Marinha No Município De São Caetano De Odivelas, Estado Do Pará, Amazônia, Brasil. Research, Society And Development, 2020, 9.7: E835974701-E835974701.

Salinas-Restrepo, Cristian, Et Al. Optimization Of The De Novo Assembly Of The Transcriptome Of The Venom Gland Of Pamphobeteus Verdolaga, Prospecting Novel Bioactive Peptides. 2020.

Sandberg, Troy E., Et Al. The Emergence Of Adaptive Laboratory Evolution As An Efficient Tool For Biological Discovery And Industrial Biotechnology. Metabolic Engineering, 2019, 56: 1-16.

Sang, Fei. Bioinformatics Analysis Of Dna Methylation Through Bisulfite Sequencing Data. In: *Dna Modifications*. Humana, New York, Ny, 2021. P. 441-450.

Santos, Ana Paula Billar Dos. Índices Químicos, Sensoriais E Microbiológicos Para Avaliação Do Frescor De Pescada Amarela (Cynoscion Acoupa) Armazenada Em Gelo. 2011. Phd Thesis. Universidade De São Paulo.

Santos, Ronilson Santos Dos. Identificação E Anotação Funcional De Proteínas Hipotéticas De Bifidobacterium Breve In Silico / Ronilson Santos Dos Santos. - 2022. 52 F. : Il. Col.

Scott, Oliver B.; Edith Chan, A. W. Scaffoldgraph: An Open-Source Library For The Generation And Analysis Of Molecular Scaffold Networks And Scaffold Trees. *Bioinformatics*, 2020, 36.12: 3930-3931.

Seemann, Torsten. Prokka: Rapid Prokaryotic Genome Annotation. *Bioinformatics*, 2014, Shumate, Alaina; Salzberg, Steven L. Liftoff: Accurate Mapping Of Gene Annotations. *Bioinformatics*, 2021, 37.12: 1639-1643.

Silva, Scheila De Avila E.; Notari, Daniel Luis; Dall'alba, Gabriel. Bioinformática: Contexto Computacional E Aplicações. Caxias Do Sul, Rs: Educs, 2020.

Simão, Felipe A., Et Al. Busco: Assessing Genome Assembly And Annotation Completeness With Single-Copy Orthologs. *Bioinformatics*, 2015, 31.19: 3210-3212.

Sohn, Jang-Il; Nam, Jin-Wu. The Present And Future Of De Novo Whole-Genome Assembly. Briefings In Bioinformatics, 2018, 19.1: 23-40.

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. Augustus: Ab Initio Prediction Of Alternative Transcripts. Nucleic Acids Res. 2006 Jul 1;34(Web Server Issue):W435-9. Doi: 10.1093/Nar/Gkl200. Pmid: 16845043; Pmcid: Pmc1538822.

Stanke, M., Schöffmann, O., Morgenstern, B., & Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, 7, 62.

Stanke, Mario, Et Al. Augustus: A Web Server For Gene Finding In Eukaryotes. *Nucleic Acids Research*, 2004.

Stanke, Mario, Et Al. Gene Prediction With A Hidden Markov Model And A New Intron Submodel. *Bioinformatics-Oxford*, 2003

Stoler, Nicholas; Nekrutenko, Anton. Sequencing Error Profiles Of Illumina Sequencing Instruments. *Nar Genomics And Bioinformatics*, 2021, 3.1: Lqab019.

Stothard, P. And Wishart, D.S. (2006) 'Automated Bacterial Genome Analysis And Annotation', *Current Opinion In Microbiology*, 9(5), Pp. 505–510. Doi:10.1016/J.Mib.2006.08.002.

Suganuma, Andréa Midori, Et Al. Metodologia Para Avaliação Da Qualidade De Pescada Amarela (Cynoscion Acoupa). 18. Siicusp: Resumos Agropecuária, 2010.

Syngai, Gareth Gordon, Et Al. Blast: An Introductory Tool For Students To Bioinformatics Applications. *Keanean Journal Of Science*, 2013.

Tang, Haibao, Et Al. An Improved Genome Release (Version Mt4. 0) For The Model Legume Medicago Truncatula. *Bmc Genomics*, 2014.

Tatusova, Tatiana, Et Al. Ncbi Prokaryotic Genome Annotation Pipeline. *Nucleic Acids Research*, 2016.

Terwilliger, T.C. *Et Al.* (1998) 'Class-Directed Structure Determination: Foundation For A Protein Structure Initiative', *Protein Science*, 7(9), Pp. 1851–1856. Doi:10.1002/Pro.5560070901.

Tørresen, Ole K., Et Al. Tandem Repeats Lead To Sequence Assembly Errors And Impose Multi-Level Challenges For Genome And Protein Databases. Nucleic Acids Research, 2019, 47.21: 10994-11006.

Trapnell, C., Pachter, L., & Salzberg, S. L. (2010). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9), 1105-1111.

Urbano, K.V. And Braga, M.B. (2016) 'Omics Technologies Applied To Prokaryotes', In *Advances In Genetics Research. Volume 16*. Nova Science Publishers Incorporated.

Vaser, Robert; Šikić, Mile. Time-And Memory-Efficient Genome Assembly With Raven. Nature Computational Science, 2021, 1.5: 332-336.

Vrancken, Bram, Et Al. Quantificando O Viés De Pré-Processamento De Amostras De Sequenciamento De Próxima Geração No Sequenciamento Completo Do Genoma Do Hiv-1. *Vírus* , 2016, 8.1:12.

Wee, Yongkiat, Et Al. The Bioinformatics Tools For The Genome Assembly And Analysis Based On Third-Generation Sequencing. *Briefings In Functional Genomics*, 2019, 18.1: 1-12.

Wick, R.R. *Et Al.* (2015) 'Bandage: Interactive Visualization Of *De Novo* Genome Assemblies', *Bioinformatics*, 31(20), Pp. 3350–3352. Doi:10.1093/Bioinformatics/Btv383.

Wojcieszer, Michał Et Al. Genomes Correction And Assembling: Present Methods And Tools. In: Photonics Applications In Astronomy, Communications, Industry, And High-Energy Physics Experiments 2014. Spie, 2014. P. 529-536.

Yandell, Mark; Ence, Daniel. A Beginner's Guide To Eukaryotic Genome Annotation. *Nature Reviews Genetics*, 2012.

Yao, Zhen, Et Al. Evaluation Of Variant Calling Tools For Large Plant Genome Re-Sequencing. *Bmc Bioinformatics*, 2020, 21.1: 1-16.

Zheng, Ancai, Et Al. Changes In Gut Microbiome Structure And Function Of Rats With Isoproterenol-Induced Heart Failure. International Heart Journal, 2019, 60.5: 1176-1183.

Zhou, C., Mccarthy, S. A., & Durbin, R. (2023). Yahs: Yet Another Hi-C Scaffolding Tool. *Bioinformatics*, *39*(1), Btac808.

Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10), 931-934.